Minimal belief revision leads to backward induction

Minimal Belief Revision leads to Backward Induction*

Andrés Perea†
Maastricht University

This Version: August 2004

Abstract

In this paper we present a model for games with perfect information in which the players, upon observing an unexpected move, may revise their beliefs about the opponents' preferences over outcomes. For a given profile P of preference relations over outcomes, we impose the following three principles: (1) players initially believe that opponents have preference relations as specified by P; (2) players believe at every instance of the game that each opponent is carrying out an optimal strategy; and (3) beliefs about the opponents' preference relations over outcomes should be revised in a minimal way. It is shown that every player whose preference relation is given by P, and who throughout the game respects common belief in the events (1), (2) and (3), has a unique optimal strategy, namely his backward induction strategy in the game induced by P. We finally show that replacing the minimal belief revision principle (3) by the more modest requirement of Bayesian updating leads exactly to the Dekel-Fudenberg procedure in the game induced by P.

Keywords: Belief revision, minimal belief change, backward induction, dynamic games.
Journal of Economic Literature Classification: C72

1. Introduction

In this paper we are concerned with the problem of how to model rationality in dynamic games. In a purely static setting, rational choice can be formalized by the requirement that players hold beliefs about the opponents' strategy choices, and choose strategies that are optimal against these beliefs. In a dynamic game, however, it may happen that a player's initial belief about the opponents' strategy choices will be contradicted by the opponents' real behavior later on in the game. At this instance, the player must revise his belief about the opponents as to explain the

*The author wishes to thank Geir Asheim, Giacomo Bonanno, Dov Samet, Robert Sugden and Shmuel Zamir for their helpful comments.

†Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: a.perea@ke.unimaas.nl, Tel: +31-43-3883922, Fax: +31-43-3884874. Web: www.personeel.unimaas.nl/a.perea

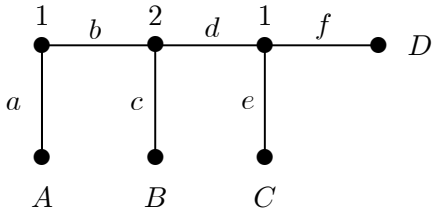

Figure 1

observed behavior. The two basic questions that we shall focus on are: How should the player revise his beliefs? and What consequences does this have for the player's own behavior?

As to illustrate the problem of belief revision, consider the game tree depicted in Figure 1. The symbols A, B, C and D denote the different outcomes, or terminal nodes, that can be reached at the end. Suppose that player 2 holds preference relation $DBCA$ over these outcomes, meaning that he strictly prefers D over B, strictly prefers B over C, and strictly prefers C over A. Assume, moreover, that (1) player 2 initially believes that player 1 has preference relation $CADB$ over outcomes, and (2) that player 2 initially believes that player 1 initially believes that player 2 will choose c. If player 2 believes, moreover, that player 1 chooses optimally given this initial belief and preference relation, he must believe at the beginning of the game that player 1 chooses a.

Suppose now that player 2 observes that player 1 has chosen action b. In this case, he must conclude that his initial belief about player 1 was wrong, and therefore needs to be replaced by a new belief that explains the event of player 1 choosing b. A possible explanation could be that player 2's initial belief about player 1's preference relation and initial belief were correct, but that player 1 has mistakenly chosen b. Although mistakes can never be ruled out in human decision making, we adopt as a *guiding principle* for our approach that players, at every possible instance of the game, believe that each of the opponents is carrying out an optimal strategy. That is, if a player observes an opponent's move that contradicts his current belief about the opponent, then he deems the event that the opponent has acted rationally more plausible than the event that the opponent has made a mistake. We shall refer to this principle as *belief in sequential rationality.*

As a consequence of this principle, player 2, upon observing b, must either revise his belief about player 1's preference relation, or revise his belief about player 1's initial belief about player 2's strategy choice, since otherwise the move b cannot be rationalized. A problem that arises here is that player 2 may choose between various belief revision procedures that rationalize the move b, and these different belief revision procedures may lead to different choices for player 2. For instance, player 2 may explain the move b by the new theory that player 1 still has preference

relation $CADB$, but that player 1 initially believes that player 2 will choose d (and not c, as player 2 believed initially). In this case, player 2 believes upon observing b that player 1 would choose e at his final decision node, and hence player 2 will choose c when adopting this belief revision procedure. Another possibility for player 2 would be to believe, upon observing b, that player 1 has preference relation $DCBA$ (instead of $CADB$, as player 2 believed initially), without revising his belief about player 1's initial belief about player 2's strategy choice. Accordingly, player 2 believes that player 1 would choose f at his final decision node, and hence player 2 will choose d when using this second belief revision procedure.

This phenomenon raises the question whether the various belief revision procedures that player 2 may adopt can be classified according to some natural criterion. A generally accepted principle in belief revision theory is that belief changes should be as small as possible, while being able to explain the newly observed information (see Schulte (2002) for an excellent discussion of the idea of minimal belief revision, and an overview of the various formalizations thereof in belief revision theory). The intuition behind this principle is that the current beliefs of a decision maker reflect, in some sense, the "best possible theory" that he can produce about the state of affairs given his current information. If these beliefs are contradicted by new observations, the decision maker therefore attempts to explain these new events by disturbing his previous beliefs as little as possible.

When applying the minimal belief revision principle to our example, it may seem, at first glance, that both belief revisions above are "equally" distant from the initial belief, as they both require one belief change: in the first belief revision, player 2 changes his belief about player 1's initial belief about player 2's choice, while leaving his belief about player 1's preference relation invariant, whereas in the second revision player 2 changes his belief about player 1's preference relation, while maintaining his belief about player 1's initial belief about player 2's choice. The problem with this argument is that the aforementioned description of player 2's beliefs only reveals a small part of the complete beliefs: player 2 does not only hold a belief about player 1's preference relation and about player 1's belief about player 2's strategy choice, but also about player 1's belief about player 2's preference relation, and about player 1's belief about player 2's belief about player 1's strategy choice, and so forth. Consequently, if player 2, in the first belief revision, changes his belief about player 1's initial belief about player 2's strategy choice from c to d, then player 2 must rationalize this belief change by changing, in addition, (1) his belief about player 1's initial belief about player 2's preference relation, or (2) his belief about player 1's initial belief about player 2's belief about player 1's choice. Namely, if player 2 initially believes that player 1 initially believes that player 2 chooses c, and player 2 moreover believes that player 1 believes in sequential rationality, as we have imposed above, then player 2 must initially believe that (1) player 1's initial belief about player 2's preference relation, together with (2) player 1's initial belief about player 2's belief about player 1's strategy choice, are such that player 1 deems c optimal for player 2. Therefore, if player 2, in the first belief revision, changes his theory by believing that player 1 initially believes that player 2 chooses d, then this must be justified by adapting the belief about player 1's initial belief about player 2's preference

relation and/or about player 2's belief about player 1's strategy choice. The bottomline is thus that the first belief revision requires, apart from the change in belief about player 1's initial belief about player 2's strategy choice, at least one more belief change. On the other hand, in the second belief revision player 2 changes his belief about player 1's preference relation, and this belief change alone is already sufficient to explain the move b by player 1. In other words, it is not necessary to complement this belief change about player 1's preference relation by an additional belief change concerning player 1. We may therefore conclude that the first belief revision requires more belief changes than the second, and the first belief revision should thus be discarded on the basis of minimal belief revision.

Since it is always possible to explain every unexpected opponent's move by a revision about the opponent's preference relation alone, the argument above implies that the principle of minimal belief revision always leads to belief revisions that concern only the opponent's preference relation, and not the opponent's belief about the other players' strategy choices. The reason, as we have seen above, is that revisions about opponent i's belief about player j's strategy choice must always be rationalized by a belief change about player i's belief about player j's preference relation or belief. The *first minimal belief revision principle* we adopt states therefore that players, when revising their beliefs about the opponents upon observing an unexpected move, should only revise their beliefs about the opponents' preference relations, and leave the other components of their beliefs unaltered.

An important implicit assumption we make when applying this first minimal belief revision principle is that all beliefs of *any order* are viewed as "equally important". That is, the belief that player i has about player j's strategy choice is considered "as important" as player i's belief about player j's belief about the other players' strategy choices. This assumption seems natural once we impose common belief in sequential rationality, as we shall do in our model, since in this case player i's belief about player j's belief about his opponents' strategies serves as a justification for player i's belief about player j's strategy choice. Common belief in sequential rationality implies, namely, that player i should believe that player j's strategy choice is optimal given player i's belief about player j's preference relation, and given player i's belief about player j's conditional beliefs about the opponents' strategy choices. Hence, player i's second order beliefs *justify* player i's first order beliefs about the opponents' strategy choices, and therefore both beliefs may be viewed as "equally important ". Similarly, common belief in sequential rationality implies that player i's k-th order beliefs justify his $(k-1)$-th order beliefs about the opponents' strategy choices for any k. For this reason, we assume that beliefs of all possible orders are viewed as "equally important" in our model, thereby justifying the first minimal belief revision principle formulated above.

The question remains whether the second belief revision described above, in which player 2 changes his belief about player 1's preference relation from $CADB$ to $DCBA$, may be regarded a minimal belief revision. As to answer this question, compare this belief revision with a third belief revision defined as follows: upon observing move b, player 2 believes that player 1's preference relation is $CDBA$, while leaving his other beliefs about player 1 invariant. Also this

belief revision procedure rationalizes the move b by player 1, and according to this new belief, player 2 expects player 1 to choose e at his final decision node, which leads to choice c for player 2 (and not d, as with the second belief revision). We argue that the third belief revision is smaller than the second, as the revised belief $CDBA$ about player 1's preference relation is "closer" to the initial belief $CADB$ than $DCBA$.

As to formalize what we mean by "closer", we measure the distance between two preference relations by counting the pairs of outcomes on which the two relations induce different pairwise rankings. For instance, $CADB$ and $DCBA$ induce different rankings on $\{A, B\}$, $\{A, D\}$ and $\{C, D\}$, and therefore the distance between both is 3. On the other hand, $CADB$ and $CDBA$ disagree solely on $\{A, B\}$ and $\{A, D\}$, meaning that the distance is only 2. Consequently, the second belief revision represents a larger belief change than the third. According to the minimal belief revision principle, the second belief revision should thus be discarded. As a *second minimal belief revision principle* we therefore require that players, when revising their belief about an opponent's preference relation, should make sure that their new belief is as close as possible to the previous belief, given the distance measure formalized above, provided that the new belief should rationalize the newly observed move(s) by this opponent.

Note that in the distance measure mentioned above, player i attaches equal weight to each pairwise ranking that player j could possibly have over outcomes. As such, it is implicitly assumed that player i is equally certain (or uncertain, if you wish) about player j's various pairwise rankings of outcomes. Of course, there are many practical examples that violate this condition, as some pairwise rankings seem intuitively less ambiguous than others, and hence belief revisions about such "less ambiguous" pairwise rankings should have a larger weight than belief revisions about "more ambiguous" pairwise rankings. The distance measure used above also assumes that there is no "correlation" between the various outcomes of the game. More precisely, it is assumed that a belief revision about an opponent's pairwise ranking of two outcomes A and B should not be a reason *per sé* to change your belief about the opponent's ranking of two other outcomes C and D. This condition may be violated in practical examples in which, intuitively, some outcomes are similar to each other. Assume, for instance, that in the example of Figure 1 it were the case that outcome A is similar to outcome C, and outcome B is similar to outcome D. As above, suppose that player 2 initially believes that player 1 has preference relation $CADB$. Then, player 2's second belief revision in which, upon observing b, he believes that player 1 has preference relation $DCBA$ should now be regarded a smaller belief change than player 2's third belief revision, in which he believes, upon observing b, that player 1 has preference relation $CDBA$. The reason is that the third belief revision contradicts the similarities of the outcomes: if player 2 believes that player 1 prefers B over A, he should also believe that player 1 prefers D over C. However, for the remainder of this paper we shall assume that players are equally certain about each of the opponents' pairwise rankings, and that there is no correlation between outcomes, and hence the distance measure introduced above makes intuitive sense.

The obvious question is now whether the third belief revision, in which player 2 changes his

belief about player 1's preference relation from $CADB$ to $CDBA$, is a minimal belief change. The answer must be "yes". Recall that player 2 initially believes that player 1 initially believes that player 2 chooses c. Hence, if player 2 changes his belief about player 1's preference relation upon observing move b, he must make sure that the new belief ranks outcome B over outcome A. However, it can be seen easily that this requires at least a distance of 2 with respect to the initial belief $CADB$, and therefore the third belief revision has minimal distance.

Although there are several minimal belief revisions for player 2 in this example, it may be verified that every minimal belief revision has the property that player 2, upon observing move b, still believes that player 1 prefers outcome C over outcome D (as player 2 believed initially). The intuition is that, in order to explain the unexpected move b by player 1, it is not necessary to change the belief about player 1's relative ranking of C and D, and hence, by minimal belief revision, player 2 should not do so. But if this is true, minimal belief revision always leads player 2 to believe, upon observing b, that player 1 would choose e at his final decision node, and hence player 2 will always choose c when acting in accordance with minimal belief revision.

Note that strategy c is the exactly the *backward induction strategy* for player 2 in the game where the players' preferences over outcomes are given by $P_1 = CADB$ and $P_2 = DBCA$, respectively. By putting $P = (P_1, P_2)$, we have thus derived the following result for this example: If player 2 has preference relation P_2, initially believes that player 1 has preference relation P_1, believes in sequential rationality and satisfies minimal belief revision, then there is a unique optimal strategy for player 2, namely his backward induction strategy in the game induced by P.

Our main theorem in this paper (Theorem 5.2) shows that a similar result is true for general games with perfect information. Consider a profile $P = (P_i)_{i \in I}$ of strict preference relations over outcomes, where P_i belongs to player i. If player i holds preference relation P_i, and respects common belief in the events that (1) players initially believe that their opponents have preference relations as given by P, (2) players believe in sequential rationality, and (3) players satisfy minimal belief revision, then player i has a unique optimal strategy, namely his backward induction strategy in the game induced by P. Here, we say that player i respects *common belief* in the event that players have a certain property if player i has this property, player i believes throughout the game that all players have this property, player i believes throughout the game that other players believe throughout the game that all players have this property, and so on.

The concepts of (common belief in) belief in sequential rationality and minimal belief revision may thus be viewed as a possible foundation for backward induction, which constitutes one of the oldest ideas in game theory. The main difference with other foundations for backward induction, such as Aumann (1995), Samet (1996), Balkenborg and Winter (1997), Stalnaker (1998), Asheim (2002) and Asheim and Perea (2004), is that in our model, players are assumed to interpret every unexpected move by an opponent as a rational move, whereas this is not the case in the latter foundations. Moreover, in our model players are allowed to revise their beliefs about the opponents' preference relations over outcomes in order to rationalize such unexpected moves, while the aforementioned foundations do not model this possibility, at least not explicitly.

Other foundations for backward induction that *do* allow players to revise their beliefs about the opponents' utilities during the game can be found in Perea (2003a, 2003b). The main difference with our approach here is that the latter two foundations use *proper belief revision,* rather than *minimal belief revision,* as a criterion to restrict the possible belief revision procedures. Proper belief revision states that whenever player i at decision node h_i revises his belief about player j, then he must not change his belief about player j's relative ranking of two strategies s_j and s'_j, if both s_j and s'_j could have led to h_i. The intuition is that such belief changes would be "unnecessary " in order to explain the event that h_i has been reached. In Section 4.2 we establish a formal relationship between minimal belief revision and proper belief revision, which proves to be important for deriving the announced theorem on backward induction.

The outline of this paper is as follows. In Section 2 we develop an epistemic model for games with perfect information that allows us to formalize statements such as "player i believes at decision node h_i that player j has preference relation P_j ", or "player i believes at decision node h_i that player j believes at decision node h_j that player k chooses strategy s_k", and so on. In this model, the relevant characteristics of a player are represented by a so-called *type,* defining a preference relation over outcomes and prescribing at every decision node a conditional belief about the opponents' strategy choices *and types.* Since types hold conditional beliefs about the opponents' types, they therefore also hold conditional beliefs about the opponents' preference relations, and about the opponents' conditional beliefs about the other players' strategy choices, and so forth. We then use this model to define the notion of common belief. In Section 3 we formalize what it means that a type "believes in sequential rationality", "satisfies minimal belief revision" and "initially believes in some profile P of preference relations". In Section 4 we derive some properties of minimal belief revision and belief in sequential rationality that are important for establishing our theorem on backward induction. In Section 5 we first show that for every profile $P = (P_i)_{i \in I}$ of preference relations, and every player i, there is at least one type for player i that respects common belief in the events that types (1) believe in sequential rationality, (2) satisfy minimal belief revision, and (3) initially believe that types hold preference relations as specified by P. We therefore guarantee that common belief in these three events is always possible. We then show that every player i type that holds preference relation P_i, and respects common belief in the three events above, has a unique optimal strategy, namely his backward induction strategy in the game induced by P. In Section 6, finally, we explore the consequences of replacing the minimal belief revision principle by the more modest requirement of Bayesian updating. It is shown that the resulting concept allows for any strategy that survives the Dekel-Fudenberg procedure in the game induced by P. Here, by the Dekel-Fudenberg procedure we mean one round of elimination of weakly dominated strategies, followed by iterative elimination of strongly dominated strategies in the game induced by P. Hence, common belief in minimal belief revision may be seen as a property that closes the gap between the Dekel-Fudenberg procedure and the concept of backward induction.

2. The Epistemic Model

2.1. Games with Perfect Information

A dynamic game is said to be with *perfect information* if every player, at each instance of the game, observes the opponents' moves that have been made until then. Formally, an *extensive form structure* $\mathcal{S}$ *with perfect information* consists of a finite game tree, a finite set I of players, for every player i a finite set H_i of decision nodes, for every decision node $h_i \in H_i$ a finite set $A(h_i)$ of available actions, and a finite set Z of terminal nodes. Perfect information is modeled by the assumption that each decision node by itself constitutes an information set. By A we denote the set of all actions, whereas H denotes the collection of all decision nodes. We assume that no chance moves occur. The definition of a strategy we shall employ coincides with the concept of a *plan of action,* as discussed in Rubinstein (1991). The difference with the usual definition is that we require a strategy only to prescribe an action at those decision nodes that the same strategy does not avoid. Formally, let $\tilde{H}_i \subseteq H_i$ be a collection of player i decision nodes, not necessarily containing all decision nodes, and let $s_i : \tilde{H}_i \to A$ be a mapping prescribing at every $h_i \in \tilde{H}_i$ some available action $s_i(h_i) \in A(h_i)$. For a given decision node $h \in H$, not necessarily belonging to player i, we say that s_i *avoids* h if there is some $h_i \in \tilde{H}_i$ on the path to h at which the prescibed action $s_i(h_i)$ deviates from the path to h. Such a mapping $s_i : \tilde{H}_i \to A$ is called a *strategy* for player i if $\tilde{H}_i$ is exactly the collection of player i decision nodes not avoided by s_i. Obviously, every strategy s_i can be obtained by first prescribing an action at all player i decision nodes, that is, constructing a strategy in the classical sense, and then deleting those player i decision nodes that are avoided by it. For a given strategy $s_i \in S_i$, we denote by $H_i(s_i)$ the collection of player i decision nodes that are not avoided by s_i. Let S_i be the set of player i strategies. For a given decision node $h \in H$ and player i, we denote by $S_i(h)$ the set of player i strategies that do not avoid h. Then, it is clear that a profile $(s_i)_{i \in I}$ of strategies reaches a decision node h if and only if $s_i \in S_i(h)$ for all players i.

2.2. Types

We shall now formally model the players in the extensive form structure $\mathcal{S}$ as decision makers under uncertainty. Our primary assumption is that every player i holds a strict, complete and transitive preference relation P_i over the set of terminal nodes, and holds at the beginning of the game, as well as at every decision node $h_i \in H_i$, a conditional belief about the opponents' strategy choices. Throughout this paper, whenever we write "preference relation over terminal nodes", we always assume that it is strict, complete and transitive. In order to keep the model as simple as possible, we assume that the conditional beliefs about the opponents' strategy choices assign at each instance of the game probability one to a single strategy choice for each of the opponents, that is, we restrict ourselves to *point-beliefs.*[1] On top of this we assume that every

[1]This assumption may be justified by the following property of games with perfect information: if a strategy s_i is optimal for player i given a *probabilistic* belief μ_i over the opponents' strategies, then there is some single

player, throughout the game, holds a conditional point-belief about the opponents' preference relations over the terminal nodes, and about the opponents' conditional beliefs about the other players' strategy choices. Moreover, each player also holds, at every instance, a conditional point-belief about the opponents' conditional beliefs concerning the other players' preferences and concerning the other players' conditional beliefs about their opponents' strategy choices, and so on. Repeating this argument inevitably leads to infinite hierarchies of conditional beliefs.

Similarly to Ben-Porath (1997), Battigalli and Siniscalchi (1999) and Perea (2004), we model such hierarchies of conditional beliefs by means of *epistemic types.* Let h_0 be the decision node that marks the beginning of the game, and let $H_i^* = H_i \cup \{h_0\}$. By applying techniques from Battigalli and Siniscalchi (1999) and Perea (2004), one can construct type spaces T_i for every player i such that every type $t_i \in T_i$ can be identified with a vector

$$(P_i(t_i), (s_j(t_i, h_i), t_j(t_i, h_i))_{h_i \in H_i^*, j \neq i}), \tag{2.1}$$

where $P_i(t_i)$ is a preference relation on the set of terminal nodes, $s_j(t_i, h_i)$ is a strategy in $S_j(h_i)$ and $t_j(t_i, h_i)$ is a type in T_j. The interpretation is that t_i holds preference relation $P_i(t_i)$, and believes at every decision node h_i that player j chooses the strategy $s_j(t_i, h_i)$ and is of type $t_j(t_i, h_i)$. Since such t_j, in turn, holds a preference relation over the terminal nodes and a conditional belief about the other players' strategy choices, every type t_i holds at every instance a conditional belief about player j's preference relation and about player j's conditional beliefs about the other players' strategy choices. In a similar fashion, one may derive from (2.1) conditional beliefs about conditional beliefs about ... about conditional beliefs, of arbitrary length. In the sequel of this paper, we often write $s_{-i}(t_i, h_i) = (s_j(t_i, h_i))_{j \neq i}$ to denote t_i's conditional belief at h_i about the opponents' strategy choices, and denote by $t_{-i}(t_i, h_i) = (t_j(t_i, h_i))_{j \neq i}$ the conditional belief of t_i at h_i about the opponents' types.

2.3. Common Belief

By $T = \cup_{i \in I} T_i$ we denote the collection of all types for all players. Let $E \subseteq T$ be some subset of types, and let t_i be a specific type for player i. We say that t_i *believes* E if $t_j(t_i, h_i) \in E$ for every opponent j and every $h_i \in H_i^*$. In words, t_i believes at every instance of the game that the opponents' types belong to E. We recursively define

$$B^1(E) = \{t \in E \mid t \text{ believes } E\}$$

and

$$B^k(E) = \{t \in B^{k-1}(E) \mid t \text{ believes } B^{k-1}(E)\}$$

strategy profile within the support of μ_i against which s_i is optimal. (Ben-Porath (1997) shows this fact in his proof of Lemma 1.2.1). Hence, every strategy choice in a game with perfect information that is justified by a probabilistic belief, can also be justified by a point-belief.

for all $k \geq 2$. Let $B^\infty(E) = \cap_{k \in \mathbb{N}} B^k(E)$. We say that t_i respects *common belief* in E if $t \in B^\infty(E)$. Hence, t_i belongs to E, believes at every instance that all opponents' types belong to E, believes at every instance that all opponents' types believe at every instance that all other players' types belong to E, and so on.

3. Belief in Sequential Rationality and Minimal Belief Revision

In this section we formalize the following three conditions: (1) a type should believe, throughout the game, that his opponents choose optimal strategies, (2) a type should revise his belief about an opponent's characteristics in a minimal way, and (3) a type should *initially* believe that the opponents' preference relations are given by some profile $P = (P_i)_{i \in I}$. We shall refer to these conditions as *belief in sequential rationality*, *minimal belief revision* and *initial belief in* P, respectively.

3.1. Belief in Sequential Rationality

Let (s_i, t_i) be a pair consisting of a strategy and a type for player i. Recall that $H_i(s_i)$ is the set of player i decision nodes that are not avoided by s_i. For a given $h_i \in H_i(s_i)$, let $z(s_i, t_i, h_i)$ denote the terminal node that is reached if the game would start at h_i, player i would choose according to s_i, and the opponents would choose according to the conditional belief $s_{-i}(t_i, h_i)$ that t_i holds at h_i about the opponents' strategy choices. We say that s_i is *sequentially rational* for t_i if for every decision node $h_i \in H_i(s_i)$ there is no strategy $s_i' \in S_i(h_i)$ such that t_i strictly prefers the terminal node $z(s_i', t_i, h_i)$ over the terminal node $z(s_i, t_i, h_i)$ with respect to his preference relation $P_i(t_i)$.

Definition 3.1. *We say that type t_i believes in sequential rationality if at every $h_i \in H_i^*$, and for every opponent j, the conditional belief $(s_j(t_i, h_i), t_j(t_i, h_i))$ about player j's strategy-type pair is such that $s_j(t_i, h_i)$ is sequentially rational for $t_j(t_i, h_i)$.*

At this stage it is important to note that not every type has a sequentially rational strategy. Consider, for instance, the extensive form structure in Figure 2. Take a type t_1 for player 1 with the preference relation $ECDBA$ over the terminal nodes. Let player 1's decision nodes be denoted by h_1^1 and h_1^2, respectively. Suppose that t_1 believes at h_1^1 that player 2 chooses the strategy (d, g), but believes at h_1^2 that player 2 chooses (d, h). The unique strategy that is optimal for t_1 at h_1^1 is (b, e). However, (b, e) is not optimal for t_1 at h_1^2, which implies that t_1 has no sequentially rational strategy. In Section 4 we shall prove that minimal belief revision and belief in sequential rationality are sufficient to imply that a type has a sequentially rational strategy. Note also that a type t_i can have at most one sequentially rational strategy.

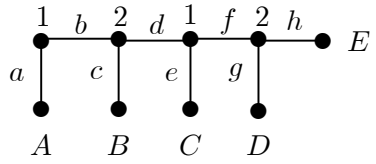

Figure 2

3.2. Minimal Belief Revision

Suppose now that a type t_i observes that decision node $h_i \in H_i$ has been reached, but cannot rationalize this event by means of his previous beliefs about player j. In this case, type t_i may be led to *revise* his belief about (1) player j's preference relation, or (2) player j's conditional beliefs about the other players' strategy choices, or both. As we have argued already in the introduction, a belief revision about player j's conditional beliefs about the other players' strategy choices must always be justified by an additional belief revision about player j's conditional beliefs about the other players' preferences and/or conditional beliefs. On the other hand, a belief revision about player j's preference relation need not be rationalized by an additional belief change. For this reason, the principle of minimal belief revision requires players to explain unexpected moves solely by belief revisions about the opponents' preference relations. Formally, let h_i^1 and h_i^2 be two decision nodes for player i such that h_i^2 follows h_i^1, and let $t_j(t_i, h_i^1)$ and $t_j(t_i, h_i^2)$ be t_i's conditional beliefs at h_i^1 and h_i^2 about player j's type. For a player j type t_j and preference relation P_j, we denote by (t_j, P_j) the type that has preference relation P_j and holds the same conditional beliefs about the opponents' strategy-type pairs as t_j. Minimal belief revision requires that $t_j(t_i, h_i^1)$ and $t_j(t_i, h_i^2)$ differ only by their preference relation, that is, $t_j(t_i, h_i^2) = (t_j(t_i, h_i^1), P_j)$ for some preference relation P_j. This requirement is formalized as condition (1) in the definition of minimal belief revision below.

The additional requirement we impose is that player i's belief revision about player j's preferences must be as small as possible. More precisely, we shall introduce a distance measure between preference relations, and require that t_i's new belief about player j's preference relation should be as close as possible to his old belief, given that the new belief should rationalize the event of reaching the decision node h_i.

Definition 3.2. *Let P^1 and P^2 be two preference relations on the set of terminal nodes Z. We define the distance $d(P^1, P^2)$ as the number of unordered pairs $\{z_1, z_2\}$ in Z at which P^1 and P^2 disagree.*

Here, we say that P^1 and P^2 disagree at $\{z_1, z_2\}$ if P^1 ranks z_1 strictly above z_2 but P^2 does not, or P^1 ranks z_1 strictly below z_2 but P^2 does not. The distance measure thus defined

coincides with the measure adopted in Ha and Haddawy (1998), and may be interpreted as a Hamming distance between preference relations, when the latter are interpreted as collections of pairwise rankings.

For a decision node $h_i \in H_i$ and type $t_j \in T_j$, we say that t_j *rationalizes the event of reaching* h_i if t_j has a sequentially rational strategy belonging to $S_j(h_i)$. The conditions (2) and (3) in the definition of minimal belief revision below state that type t_i, upon reaching decision node h_i, should change his belief about player j's preference relation in a minimal way, provided that his new belief about player j's type rationalizes the event of reaching h_i.

We are now ready to define minimal belief revision. Let t_i be a type and let $h_i^1, h_i^2 \in H_i^*$ such that h_i^2 follows h_i^1 and no other $h_i \in H_i$ lies between h_i^1 and h_i^2.

Definition 3.3. *We say that t_i satisfies minimal belief revision at h_i^2 if for every opponent j there is some preference relation P_j^2 such that*
(1) $t_j(t_i, h_i^2) = (t_j(t_i, h_i^1), P_j^2)$,
(2) $t_j(t_i, h_i^2)$ rationalizes the event of reaching h_i^2, and
(3) there is no other preference relation $\tilde{P}_j^2$ such that $(t_j(t_i, h_i^1), \tilde{P}_j^2)$ rationalizes the event of reaching h_i^2, and $d(P_j(t_j(t_i, h_i^1)), \tilde{P}_j^2) < d(P_j(t_j(t_i, h_i^1)), P_j^2)$.

We finally say that t_i satisfies minimal belief revision if it does so at every decision node h_i^2.

3.3. Initial Belief in P

Let $P = (P_i)_{i \in I}$ be some profile of preference relations.

Definition 3.4. *We say that type t_i initially believes in P if $P_j(t_j(t_i, h_0)) = P_j$ for all opponents j.*

Here, $t_j(t_i, h_0)$ is t_i's initial belief about player j's type, and $P_j(t_j(t_i, h_0))$ thus reflects t_i's initial belief about player j's preference relation. Note, however, that t_i may change his belief about j's preference relation if the game moves from h_0 to some other decision node h_i.

4. Properties of Minimal Belief Revision

As a preparatory step towards our backward induction theorem, we first derive some properties of minimal belief revision that will be applied in Section 5 for showing the announced relationship with backward induction.

4.1. Existence of Sequentially Rational Strategies

In the example of Figure 1 we have seen that not every type has a sequentially rational strategy. Namely, if the type t_1 believes at his first decision node h_1^1 that player 2 chooses (d, g), but

believes at his second decision node h_1^2 that player 2 chooses (d, h), then t_1 has no sequentially rational strategy. The reason for this is that t_1's conditional beliefs at h_1^2 contradict Bayesian updating: t_1's beliefs at h_1^1 about player 2's behavior are compatible with the event of reaching h_1^2, and therefore Bayesian updating implies that t_1's beliefs at h_1^2 should coincide with his beliefs at h_1^1.

We shall now provide a formalization of the above mentioned Bayesian updating requirement and show in Lemma 4.2 that it guarantees the existence of a sequentially rational strategy. Let h_i^1 and h_i^2 be two decision nodes in H_i^* such that h_i^2 follows h_i^1, and there is no player i decision node between h_i^1 and h_i^2.

Definition 4.1. *We say that t_i satisfies Bayesian updating at h_i^2 if for every opponent j for which $s_j(t_i, h_i^1) \in S_j(h_i^2)$, it holds that $s_j(t_i, h_i^2) = s_j(t_i, h_i^1)$.*

In other words, if t_i's belief at h_i^1 about player j's strategy choice does not contradict the event of reaching h_i^2, then t_i should maintain at h_i^2 his previous belief about player j's strategy choice. We say that t_i satisfies Bayesian updating if it does so at every decision node.

Lemma 4.2. *Every type that satisfies Bayesian updating has a sequentially rational strategy.*

The proof of this lemma is based on Theorem 3.1 in Perea (2002). The 'if' part of this theorem states that if an "updating system" satisfies "updating consistency", then every "locally sequentially rational" strategy is sequentially rational. In order to state the 'if ' part of this theorem more precisely, we must first formally define the terms "updating system", "updating consistency" and "locally sequentially rational strategy". As to simplify matters we shall define these objects directly within our special context of games with perfect information, and restrict ourselves to "updating systems" that always assign probability one to one particular strategy choice for every opponent. The reason for the latter is that the conditional belief vectors in our epistemic model always assign probability one to one particular strategy choice for each opponent.

An *updating system* for player i is a vector $c_i = (c_i(h_i))_{h_i \in H_i}$ where $c_i(h_i) \in S_{-i}(h_i)$ for every decision node $h_i \in H_i$. Here, $S_{-i}(h_i) = \times_{j \neq i} S_j(h_i)$, and $c_i(h_i)$ represents player i's conditional belief at h_i about the opponents' strategy choices. For a given decision node h_i and conditional beliefs $c_i(h_i), c_i'(h_i) \in S_{-i}(h_i)$, we say that $c_i(h_i)$ and $c_i'(h_i)$ are *equivalent at h_i* if for every strategy $s_i \in S_i(h_i)$ it holds that the strategy profiles $(s_i, c_i(h_i))$ and $(s_i, c_i'(h_i))$ lead to the same terminal node. Hence, $c_i(h_i)$ and $c_i'(h_i)$ only differ at decision nodes that do not precede nor follow h_i. The updating system c_i is called *updating consistent* if for every two decision nodes h_i^1 and h_i^2 where h_i^1 precedes h_i^2 and $c_i(h_i^1) \in S_{-i}(h_i^2)$, it holds that $c_i(h_i^2)$ and $c_i(h_i^1)$ are equivalent at h_i^2. An *extended strategy* for player i is a vector $\tilde{s}_i = (\tilde{s}_i(h_i))_{h_i \in H_i}$ where $\tilde{s}_i(h_i) \in A(h_i)$ for every decision node h_i. The difference with a strategy as defined in this paper is thus that an extended strategy also prescribes actions at decision nodes that are avoided by it, whereas a strategy does not. An extended strategy $\tilde{s}_i$ is called *locally sequentially rational* with respect to

an updating system c_i and a preference relation P_i over the terminal nodes if at every decision node h_i the action $\tilde{s}_i(h_i)$ is optimal against the actions prescribed by $c_i(h_i)$ *and* $\tilde{s}_i$ in the subgame that follows h_i. We say that a (non-extended) strategy s_i is locally sequentially rational with respect to c_i and P_i if there is an extended strategy $\tilde{s}_i$ such that $\tilde{s}_i$ is locally sequentially rational with respect to c_i and P_i, and $\tilde{s}_i$ coincides with s_i at decision nodes in $H_i(s_i)$. Finally, a strategy s_i is called *sequentially rational* with respect to c_i and P_i if at every decision node $h_i \in H_i(s_i)$ there is no other strategy $s_i' \in S_i(h_i)$ such that the terminal node reached by s_i' and $c_i(h_i)$ is strictly preferred by P_i over the terminal node reached by s_i and $c_i(h_i)$. The 'if' part of Theorem 3.1 in Perea (2002), when applied to our specific context, can now be stated as follows.

Lemma 4.3. *Let c_i be an updating system that is updating consistent, and let P_i be a preference relation over the terminal nodes. Then, every strategy that is locally sequentially rational with respect to c_i and P_i is also sequentially rational with respect to c_i and P_i.*

We are now ready to prove Lemma 4.2.

Proof of Lemma 4.2. Let t_i be a type with preference relation P_i that satisfies Bayesian updating. We show that t_i has a sequentially rational strategy. For every decision node $h_i \in H_i$ define $c_i(h_i) = s_{-i}(t_i, h_i))$, which is an element in $S_{-i}(h_i)$. Hence, the vector $c_i = (c_i(h_i))_{h_i \in H_i}$ is an updating system. Since t_i satisfies Bayesian updating, it immediately follows that the updating system c_i is updating consistent. By Lemma 4.3 we then know that every locally sequentially rational strategy with respect to c_i and P_i is sequentially rational with respect to c_i and P_i. Since it is clear that every sequentially rational strategy with respect to c_i and P_i is also sequentially rational for t_i, it suffices to show that there is a locally sequentially rational strategy with respect to c_i and P_i.

By a simple backward induction procedure, one can define for every player i decision node $h_i \in H_i$ some action $a(h_i)$ such that every action $a(h_i)$ is optimal with respect to P_i against (1) the actions prescribed by $c_i(h_i)$ at the opponents' decision nodes following h_i, and (2) his own actions $a(h_i')$ at decision nodes $h_i' \in H_i$ following h_i. Then, by construction of the actions $a(h_i)$, the extended strategy $\tilde{s}_i = (a(h_i))_{h_i \in H_i}$ is locally sequentially rational with respect to c_i and P_i. Let s_i be the unique strategy that coincides with $\tilde{s}_i$ at all decision nodes in $H_i(s_i)$. Hence, s_i is locally sequentially rational with respect to c_i and P_i. As we have seen above, this implies that s_i is sequentially rational with respect to c_i and P_i. But then, s_i is sequentially rational for t_i. This completes the proof of this lemma. ■

We shall now prove that minimal belief revision and belief in sequential rationality lead to Bayesian updating.

Lemma 4.4. *Let t_i be a type that believes in sequential rationality and satisfies minimal belief revision. Then, t_i satisfies Bayesian updating.*

Proof. Choose a type t_i that believes in sequential rationality and satisfies minimal belief revision. Let h_i^1, h_i^2 be two decision nodes in H_i^* such that h_i^2 follows h_i^1, and no player i decision node is between h_i^1 and h_i^2. Let j be an opponent for which $s_j(t_i, h_i^1)$ belongs to $S_j(h_i^2)$. As t_i believes in sequential rationality, it must be the case that $s_j(t_i, h_i^1)$ is sequentially rational for type $t_j(t_i, h_i^1)$. The fact that $s_j(t_i, h_i^1) \in S_j(h_i^2)$ then implies that the type $t_j(t_i, h_i^1)$ itself already rationalizes the event of reaching h_i^2. By minimal belief revision, it must therefore be the case that $t_j(t_i, h_i^2) = t_j(t_i, h_i^1)$. Since t_i believes in sequential rationality, and since $s_j(t_i, h_i^1)$ is the unique sequentially rational strategy for $t_j(t_i, h_i^1)$, it follows that $s_j(t_i, h_i^2) = s_j(t_i, h_i^1)$, which implies that t_i satisfies Bayesian updating. This completes the proof. ■

By combining Lemma 4.2 and Lemma 4.4, we obtain the following corollary.

Corollary 4.5. *Let t_i be a type that believes in sequential rationality and satisfies minimal belief revision. Then, t_i has a sequentially rational strategy.*

4.2. Relation with Proper Belief Revision

We next prove that minimal belief revision and belief in Bayesian updating leads to *proper belief revision*: a concept that has been put forward in Perea (2003a, 2003b and 2004). This result will prove to be crucial for establishing the announced relationship with backward induction. Informally, proper belief revision states that a player who wishes to revise his beliefs at decision node h about opponent j's preference relation, should not change his belief about the opponent's relative ranking of two strategies s_j and s_j' if both s_j and s_j' could have led to h. The intuition is that the player, upon arriving at h, cannot exclude any of the opponent's strategies s_j and s_j', and therefore there is no reason for him to change his belief about the opponent's relative ranking of s_j and s_j'. In order to introduce proper belief revision formally, we need some more notation and definitions. Let t_i be a type for player i, and $h_i \in H_i^*$ some decision node. For a given strategy $s_i \in S_i(h_i)$, recall that $z(s_i, t_i, h_i)$ denotes the terminal node that would be reached if the game would start at h_i, player i would choose according to s_i, and player i's opponents would choose according to $s_{-i}(t_i, h_i)$. For two strategies $s_i, s_i' \in S_i(h_i)$,we say that t_i strictly prefers strategy s_i over strategy s_i' at decision node h_i if t_i strictly prefers the terminal node $z(s_i, t_i, h_i)$ over the terminal node $z(s_i', t_i, h_i)$. Now, let t_i be a type for player i, let $j \neq i$ be an opponent, let h_i and h_j be decision nodes for players i and j, respectively, and let s_j, s_j' be two player j strategies in $S_j(h_j)$.

Definition 4.6. *We say that t_i believes at h_i that player j at h_j strictly prefers strategy s_j over strategy s_j' if type $t_j(t_i, h_i)$ strictly prefers s_j over s_j' at h_j.*

Now, let t_i be a type for player i, and let h_i^1, h_i^2 be two decision nodes in H_i^* such that h_i^2 follows h_i^1 and no other player i decision node is between h_i^1 and h_i^2.

Definition 4.7. *We say that t_i satisfies proper belief revision at h_i^2 if for every opponent j, every decision node $h_j \in H_j$ and every two strategies s_j, s_j' that belong to both $S_j(h_j)$ and $S_j(h_i^2)$ the following holds: t_i believes at h_i^2 that player j at h_j strictly prefers s_j over s_j' if and only if t_i believes so at h_i^1.*

Note that $s_j, s_j' \in S_j(h_i^2)$ implies that both s_j and s_j' could have led to h_i^2. We say that type t_i satisfies proper belief revision if t_i does so at each of his decision nodes.

Before showing that minimal belief revision and belief in Bayesian updating imply proper belief revision, we prove the following lemma. It states that the distance between two preference relations P^1 and P^2 can be reduced strictly by applying the following procedure: First, take an unordered pair $\{a, b\}$ of terminal nodes on which P^1 and P^2 disagree, and then interchange the roles of a and b in P^2 without changing the roles of the other nodes.

Lemma 4.8. *Let P^1 and P^2 be two preference relations on the set Z of terminal nodes, and let $\{a, b\}$ be an unordered pair of terminal nodes on which P^1 and P^2 disagree. Let u^2 be an arbitrary utility representation of P^2, and let the utility function $\tilde{u}^2$ be given by*

$$\tilde{u}^2(z) = \begin{cases} u^2(b), & \text{if } z = a, \\ u^2(a), & \text{if } z = b, \\ u^2(z), & \text{otherwise.} \end{cases}$$

Let $\tilde{P}^2$ be the preference relation induced by $\tilde{u}_2$. Then, $d(P^1, \tilde{P}^2) < d(P^1, P^2)$.

The proof can be found in the appendix. We are now able to prove the following result.

Theorem 4.9. *Let t_i be a type that satisfies minimal belief revision and believes that every opponent satisfies Bayesian updating. Then, t_i satisfies proper belief revision.*

Proof. For a given type $t_i \in T_i$, decision node $h_i \in H_i^*$, and strategy $s_i \in S_i(h_i)$, recall that $z(s_i, t_i, h_i)$ is the terminal node that is reached if the game would start at h_i, player i chooses s_i and i's opponents would act according to $s_{-i}(t_i, h_i)$. Let

$$Z(t_i, h_i) = \{z(s_i, t_i, h_i) \mid s_i \in S_i(h_i)\}$$

be the set of terminal nodes that can be reached if the game would start at h_i and player i's opponents would act according to $s_{-i}(t_i, h_i)$.

Let t_i be a type for player i that satisfies minimal belief revision and believes that every opponent satisfies Bayesian updating. We prove that t_i satisfies proper belief revision. Suppose, contrary to what we want to prove, that t_i does not satisfy proper belief revision. Then, there must be two decision nodes $h_i^1, h_i^2 \in H_i^*$ such that h_i^2 follows h_i^1 and no other player i decision node is between h_i^1 and h_i^2, an opponent j, a decision node $h_j^* \in H_j$ and two strategies $s_j, s_j' \in$

$S_j(h_j^*) \cap S_j(h_i^2)$ such that: t_i believes at h_i^1 that player j strictly prefers s_j over s_j' at h_j^*, but does not believe so at h_i^2.[2] Let $t_j^1 = t_j(t_i, h_i^1)$ and $t_j^2 = t_j(t_i, h_i^2)$, and let P_j^1 and P_j^2 denote the preference relations of t_j^1 and t_j^2, respectively. Since t_i satisfies minimal belief revision, it must be the case that $t_j^2 = (t_j^1, P_j^2)$. In particular, t_j^1 and t_j^2 hold the same conditional belief at h_j^* about the opponents' strategy choices, that is, $s_{-j}(t_j^1, h_j^*) = s_{-j}(t_j^2, h_j^*)$.

Since t_i believes at h_i^1 that player j strictly prefers s_j over s_j' at h_j^*, but does not believe so at h_i^2, we may conclude that P_j^1 strictly prefers $z(s_j, t_j^1, h_j^*)$ over $z(s_j', t_j^1, h_j^*)$, but P_j^2 strictly prefers $z(s_j', t_j^1, h_j^*)$ over $z(s_j, t_j^1, h_j^*)$. Let u_j^2 be some arbitrary utility representation of P_j^2, and let the utility function $\tilde{u}_j^2$ be given by

$$\tilde{u}_j^2(z) = \begin{cases} u_j^2(z(s_j', t_j^1, h_j^*)), & \text{if } z = z(s_j, t_j^1, h_j^*), \\ u_j^2(z(s_j, t_j^1, h_j^*)), & \text{if } z = z(s_j', t_j^1, h_j^*), \\ u_j^2(z), & \text{otherwise.} \end{cases} \tag{4.1}$$

Let $\tilde{P}_j^2$ be the preference relation induced by $\tilde{u}_j^2$. Since P_j^1 and P_j^2 disagree on $\{z(s_j, t_j^1, h_j^*), z(s_j', t_j^1, h_j^*)\}$, we know by Lemma 4.8 that $d(P_j^1, \tilde{P}_j^2) < d(P_j^1, P_j^2)$.

We now prove that the type $\tilde{t}_j^2 = (t_j^1, \tilde{P}_j^2)$ rationalizes the event of reaching h_i^2, which would contradict our assumption that t_i satisfies minimal belief revision. Since $t_j^1 = t_j(t_i, h_i^1)$, and t_i believes that player j satisfies Bayesian updating, it follows that t_j^1 satisfies Bayesian updating. Since $t_j^2 = (t_j^1, P_j^2)$ and $\tilde{t}_j^2 = (t_j^1, \tilde{P}_j^2)$, we have that also t_j^2 and $\tilde{t}_j^2$ satisfy Bayesian updating. By Lemma 4.2 we know that t_j^2 and $\tilde{t}_j^2$ have a sequentially rational strategy, which must then be unique. Let s_j^2 and $\tilde{s}_j^2$ be the unique sequentially rational strategies for types t_j^2 and $\tilde{t}_j^2$, respectively. Recall that, by definition, $t_j^2 = t_j(t_i, h_i^2)$. As t_i satisfies minimal belief revision, t_j^2 must rationalize the event of reaching h_i^2 and hence $s_j^2 \in S_j(h_i^2)$. In order to prove that $\tilde{t}_j^2$ rationalizes the event of reaching h_i^2, we must show that $\tilde{s}_j^2 \in S_j(h_i^2)$.

For every $h_j \in H_j$ preceding h_i^2, let $a(h_j, h_i^2)$ be the unique action at h_j leading to h_i^2. In order to show that $\tilde{s}_j^2 \in S_j(h_i^2)$, we prove that $\tilde{s}_j^2(h_j) = a(h_j, h_i^2)$ for all $h_j \in H_j(\tilde{s}_j^2)$ preceding h_i^2. Choose some $h_j \in H_j(\tilde{s}_j^2)$ preceding h_i^2. As $s_j^2 \in S_j(h_i^2)$, we have that $h_j \in H_j(s_j^2)$ and $s_j^2(h_j) = a(h_j, h_i^2)$. By assumption, s_j^2 is sequentially rational for $t_j^2 = (t_j^1, P_j^2)$, which means in particular that s_j^2 is optimal for t_j^2 at h_j. Hence, P_j^2 strictly prefers $z(s_j^2, t_j^1, h_j)$ over all other nodes in $Z(t_j^1, h_j)$. We distinguish two cases.

Case 1. Suppose that $z(s_j^2, t_j^1, h_j) \neq z(s_j', t_j^1, h_j^*)$, where s_j' is the strategy as discussed above. Since P_j^2 strictly prefers $z(s_j^2, t_j^1, h_j)$ over all other nodes in $Z(t_j^1, h_j)$, we have by (4.1) that $\tilde{P}_j^2$ also strictly prefers $z(s_j^2, t_j^1, h_j)$ over all other nodes in $Z(t_j^1, h_j)$. This implies that s_j^2 is optimal for

[2] Note that if t_i believes at h_i^1 that player j is indifferent at h_j^* between s_j and s_j', then necessarily $z(s_j, t_j(t_i, h_i^1), h_j^*) = z(s_j', t_j(t_i, h_i^1), h_j^*)$. By minimal belief revision of t_i, we have that $t_j(t_i, h_i^1)$ and $t_j(t_i, h_i^2)$ hold the same conditional beliefs, and hence $z(s_j, t_j(t_i, h_i^2), h_j^*) = z(s_j', t_j(t_i, h_i^2), h_j^*)$, which implies that t_i believes at h_i^2 that player j is indifferent between s_j and s_j'.

$\tilde{t}_j^2$ at h_j. Since we know that $\tilde{s}_j^2$ is optimal for $\tilde{t}_j^2$ at h_j, it follows that $\tilde{s}_j^2(h_j) = s_j^2(h_j) = a(h_j, h_i^2)$, which was to show.

Case 2. Suppose that $z(s_j^2, t_j^1, h_j) = z(s_j', t_j^1, h_j^*)$. In this case, the terminal node $z(s_j^2, t_j^1, h_j)$ follows both h_j and h_j^*. Hence, it must be the case that h_j precedes or follows h_j^*. We distinguish two subcases.

Case 2.1. Suppose that h_j precedes h_j^*. Since $z(s_j^2, t_j^1, h_j)$ follows h_j^*, it must be the case that $s_{-j}(t_j^1, h_j) \in S_{-j}(h_j^*)$. We have seen above that t_j^1 satisfies Bayesian updating, which then implies that $s_{-j}(t_j^1, h_j^*) = s_{-j}(t_j^1, h_j)$. As $s_j \in S_j(h_j^*)$, it follows that $s_j \in S_j(h_j)$ and that $z(s_j, t_j^1, h_j) = z(s_j, t_j^1, h_j^*)$. Since P_j^2 strictly prefers $z(s_j^2, t_j^1, h_j) = z(s_j', t_j^1, h_j^*)$ over all other nodes in $Z(t_j^1, h_j)$, it follows by (4.1) that $\tilde{P}_j^2$ strictly prefers $z(s_j, t_j^1, h_j) = z(s_j, t_j^1, h_j^*)$ over all other nodes in $Z(t_j^1, h_j)$. Hence, s_j is optimal for $\tilde{t}_j^2$ at h_j. Since, by assumption, $\tilde{s}_j^2$ is optimal for $\tilde{t}_j^2$ at h_j, it follows that $\tilde{s}_j^2(h_j) = s_j(h_j)$. Since $s_j \in S_j(h_i^2)$, we have that $s_j(h_j) = a(h_j, h_i^2)$. Hence, $\tilde{s}_j^2(h_j) = a(h_j, h_i^2)$, which was to show.

Case 2.2. Suppose that h_j^* precedes h_j. As $z(s_j', t_j^1, h_j^*) = z(s_j^2, t_j^1, h_j)$ follows h_j, we must have that $s_{-j}(t_j^1, h_j^*) \in S_{-j}(h_j)$. By Bayesian updating of t_j^1, we may then conclude that $s_{-j}(t_j^1, h_j) = s_{-j}(t_j^1, h_j^*)$. Since $s_j \in S_j(h_i^2)$ and h_j precedes h_i^2, we have that $s_j \in S_j(h_j)$ as well. Combined with the fact that $s_{-j}(t_j^1, h_j) = s_{-j}(t_j^1, h_j^*)$, this implies that $z(s_j, t_j^1, h_j) = z(s_j, t_j^1, h_j^*)$. Since P_j^2 strictly prefers $z(s_j^2, t_j^1, h_j) = z(s_j', t_j^1, h_j^*)$ over all other nodes in $Z(t_j^1, h_j)$, it follows by (4.1) that $\tilde{P}_j^2$ strictly prefers $z(s_j, t_j^1, h_j) = z(s_j, t_j^1, h_j^*)$ over all other nodes in $Z(t_j^1, h_j)$. We may thus conclude that s_j is optimal for $\tilde{t}_j^2$ at h_j. As $\tilde{s}_j^2$ is optimal for $\tilde{t}_j^2$ at h_j as well, it follows that $\tilde{s}_j^2(h_j) = s_j(h_j)$. By assumption, $s_j \in S_j(h_i^2)$, implying that $s_j(h_j) = a(h_j, h_i^2)$. Hence, we may conclude that $\tilde{s}_j^2(h_j) = a(h_j, h_i^2)$, which was to show.

From Case 1 and 2 we may therefore conclude that $\tilde{s}_j^2(h_j) = a(h_j, h_i^2)$ for all decision nodes $h_j \in H_j(\tilde{s}_j^2)$ preceding h_i^2. This, in turn, implies that $\tilde{s}_j^2 \in S_j(h_i^2)$. As $\tilde{s}_j^2$ is the unique sequentially rational strategy for $\tilde{t}_j^2$, this leads to the conclusion that $\tilde{t}_j^2 = (t_j^1, \tilde{P}_j^2)$ rationalizes the event of reaching h_i^2. Since we have seen that $d(P_j^1, \tilde{P}_j^2) < d(P_j^1, P_j^2)$, we have thus found a preference relation $\tilde{P}_j^2$ with the properties that $(t_j^1, \tilde{P}_j^2)$ rationalizes the event of reaching h_i^2, but $d(P_j^1, \tilde{P}_j^2) < d(P_j^1, P_j^2)$. This, however, contradicts our assumption that t_i satisfies minimal belief revision. Therefore, the assumption that t_i does not satisfy proper belief revision cannot be true. Hence, t_i must satisfy proper belief revision. This completes the proof of our theorem. ■

5. Relation with Backward Induction

In this section we show that common belief in the events that types (1) believe in sequential rationality, (2) satisfy minimal belief revision and (3) initially believe in some profile $P = (P_i)_{i \in I}$ of preference relations, leads to backward induction in the game induced by P. We divide this result into two parts. In the first part, Theorem 5.1, it is shown that for every player there is at

least one type that respects common belief in the three events listed above. As such, common belief in these three events is always possible. The second part, Theorem 5.2, shows that every type t_i that has preference relation P_i and satisfies common belief in the events that types believe in sequential rationality, satisfy minimal belief revision and initially believe in P, must choose his backward induction strategy in the game induced by P.

For the proof of Theorem 5.1 and the statement of Theorem 5.2, we need the following definitions. Let $\mathcal{S}$ be an extensive form structure with perfect information, and $P = (P_i)_{i \in I}$ a profile of preference relations on the set of terminal nodes. Then, the pair $(\mathcal{S}, P)$ may be interpreted as a *game,* and the backward induction procedure in the game $(\mathcal{S}, P)$ leads to a unique backward induction action $a^*(h_i)$ at every decision node h_i. For every player i, let s_i^* be the unique strategy that chooses the backward induction action $a^*(h_i)$ at every $h_i \in H_i(s_i^*)$. We refer to s_i^* as the *backward induction strategy* for player i in $(\mathcal{S}, P)$.

Theorem 5.1. *Let $\mathcal{S}$ be an extensive form structure with perfect information, and $P = (P_i)_{i \in I}$ a profile of preference relations on the set of terminal nodes. Then, for every player i there is a type t_i that respects common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in P.*

Proof. For every player i, decision node $h_i \in H_i^*$ and opponent $j \neq i$, let $s_j^*(h_i)$ be the unique strategy for player j with the following properties: (1) at every decision node $h_j \in H_j(s_j^*(h_i))$ preceding h_i, the strategy $s_j^*(h_i)$ prescribes the unique action that leads to h_i, and (2) at every decision node $h_j \in H_j(s_j^*(h_i))$ not preceding h_i, it prescribes the backward induction action $a^*(h_j)$ in the game $(\mathcal{S}, P)$. Then, by construction, $s_j^*(h_i)$ is a strategy in $S_j(h_i)$. Moreover, $s_j^*(h_0)$ coincides with the backward induction strategy s_j^* in $(\mathcal{S}, P)$.

For every player i, denote by β_i the conditional belief vector about the opponents' strategy choices in which player i, at every decision node $h_i \in H_i^*$, believes that each opponent j chooses the strategy $s_j^*(h_i) \in S_j(h_i)$. By construction, the unique strategy that is sequentially rational for player i with respect to the conditional belief vector β_i and the preference relation P_i is his backward induction strategy s_i^* in $(\mathcal{S}, P)$.

Fix a player i and an opponent $j \neq i$. For every decision node $h_i \in H_i^*$ we shall define a conditional belief $P_j(h_i)$ for player i about player j's preference relation. We proceed recursively, starting from h_0. At h_0, let $P_j(h_0) = P_j$. Now, take a decision node $h_i^2 \in H_i^*$ and suppose that $P_j(h_i^1)$ has already been defined for all $h_i^1 \in H_i^*$ that precede h_i^2. Let h_i^1 be the unique decision node in H_i^* preceding h_i^2 with the property that no other player i decision node lies between h_i^1 and h_i^2. By assumption, $P_j(h_i^1)$ has already been defined. We can now choose a preference relation $P_j(h_i^2)$ with the following properties: (1) the conditional belief vector β_j for player j and the preference relation $P_j(h_i^2)$ together rationalize the event of reaching h_i^2, and (2) there is no preference relation $\tilde{P}_j(h_i^2)$ that together with β_j rationalizes the event of reaching h_i^2, and for which $d(P_j(h_i^1), \tilde{P}_j(h_i^2)) < d(P_j(h_i^1), P_j(h_i^2))$. In this way, a conditional belief $P_j(h_i)$ about player

j's preference relation can be defined for every player i, every opponent j, and every decision node $h_i \in H_i^*$.

We may now construct a set of types

$$T^* = \{t_j(h_i) \mid i, j \in I,\ i \neq j \text{ and } h_i \in H_i^*\}$$

with the following properties:
(1) the preference relation for $t_j(h_i)$ is equal to $P_j(h_i)$;
(2) the conditional belief vector of $t_j(h_i)$ about the opponents' strategy choices is given by β_j, that is, $s_k(t_j(h_i), h_j) = s_k^*(h_j)$ for all $h_j \in H_j^*$ and all opponents $k \neq j$;
(3) the conditional belief of $t_j(h_i)$ at decision node $h_j \in H_j^*$ about opponent k's type is equal to $t_k(h_j)$.

We now prove that every type $t_j(h_i) \in T^*$ respects common belief in the event that types believe in sequential rationality, satisfy minimal belief revision and initially believe in P. By construction, every type $t \in T^*$ believes, at each of his decision nodes, that each of his opponents' types belongs to T^*. It is therefore sufficient to show that every type $t_j(h_i) \in T^*$ believes in sequential rationality, satisfies minimal belief revision, and initially believes in P.

Initial belief in P. Choose an arbitrary type $t_j(h_i) \in T^*$. By definition, $t_j(h_i)$ believes at h_0 that every opponent k is of type $t_k(h_0)$. Since $t_k(h_0)$ has preference relation $P_k(h_0)$ and since, by construction, $P_k(h_0) = P_k$, we have that $t_j(h_i)$ believes at h_0 that every opponent k has preference relation P_k. Hence, $t_j(h_i)$ initially believes in P.

Minimal belief revision. Choose an arbitrary type $t_j(h_i) \in T^*$ and some opponent $k \neq j$. Take some decision nodes h_j^1 and h_j^2 such that h_j^2 follows h_j^1 and no other player j decision node is between h_j^1 and h_j^2. By definition, $t_j(h_i)$ believes at h_j^1 that player k has type $t_k(h_j^1)$, and believes at h_j^2 that player k has type $t_k(h_j^2)$. By construction of $t_k(h_j^1)$ and $t_k(h_j^2)$ we know that $t_k(h_j^1)$ has preference relation $P_k(h_j^1)$, that $t_k(h_j^2)$ has preference relation $P_k(h_j^2)$, and that $t_k(h_j^1)$ and $t_k(h_j^2)$ have identical conditional beliefs about the opponents' strategies and types. As such,

$$t_k(h_j^2) = (t_k(h_j^1), P_k(h_j^2)).$$

Moreover, by construction of the preference relation $P_k(h_j^2)$, we know that (1) the conditional belief vector β_k for player k and the preference relation $P_k(h_j^2)$ together rationalize the event of reaching h_j^2, and (2) there is no preference relation $\tilde{P}_k(h_j^2)$ that together with β_k rationalizes the event of reaching h_j^2, and for which $d(P_k(h_j^1), \tilde{P}_k(h_j^2)) < d(P_k(h_j^1), P_k(h_j^2))$. As β_k is the conditional belief vector for types $t_k(h_j^1)$ and $t_k(h_j^2)$ about the opponents' strategies, it follows from (1) and (2) above that $t_j(h_i)$ satisfies minimal belief revision.

Belief in sequential rationality. Take some arbitrary $t_j(h_i) \in T^*$, a decision node $h_j \in H_j^*$ and some opponent k. By definition, $t_j(h_i)$ believes at h_j that opponent k has type $t_k(h_j)$ and chooses strategy $s_k^*(h_j)$. We prove that $s_k^*(h_j)$ is sequentially rational for $t_k(h_j)$. We do so by induction on the number of decision nodes in H_j^* that precede h_j.

Assume first that h_j is not preceded by any decision node in H_j^*, that is, $h_j = h_0$. In this case, $t_k(h_0)$ has preference relation $P_k(h_0)$ which, by construction, is equal to P_k. Since $t_k(h_0)$'s conditional belief vector about the opponents' strategy choices is given by β_k, it follows that $t_k(h_0)$ has a unique sequentially rational strategy, namely his backward induction strategy in $(\mathcal{S}, P)$, which is $s_k^* = s_k^*(h_0)$. We thus have that $s_k^*(h_0)$ is sequentially rational for $t_k(h_0)$, which was to show.

Now, take some decision node $h_j^2 \in H_j^* \backslash \{h_0\}$ and assume that for every $h_j^1 \in H_j^*$ preceding h_j^2 it holds that $s_k^*(h_j^1)$ is sequentially rational for $t_k(h_j^1)$. We prove that $s_k^*(h_j^2)$ is sequentially rational for $t_k(h_j^2)$. Hence, we must prove for every $h_k \in H_k(s_k^*(h_j^2))$ that $s_k^*(h_j^2)$ is optimal for $t_k(h_j^2)$ at h_k. We distinguish two cases.

Case 1. Assume that $h_k \in H_k(s_k^*(h_j^2))$ and that h_k does not precede h_j^2. Then, by definition of $s_k^*(h_j^2)$, we have that $s_k^*(h_j^2)$ prescribes the backward induction action $a^*(h_k')$ at every player k decision node h_k' weakly following h_k. Suppose, contrary to what we want to prove, that $s_k^*(h_j^2)$ is not optimal for $t_k(h_j^2)$ at h_k. Hence, there is some $s_k(h_j^2) \in S_k(h_k)$ such that $t_k(h_j^2)$ strictly prefers $s_k(h_j^2)$ over $s_k^*(h_j^2)$ at h_k. Now, let the strategy $\tilde{s}_k(h_j^2)$ be such that (1) $\tilde{s}_k(h_j^2)$ coincides with $s_k(h_j^2)$ at all decision nodes in $H_k(\tilde{s}_k(h_j^2))$ weakly following h_k, and (2) $\tilde{s}_k(h_j^2)$ coincides with $s_k^*(h_j^2)$ at all other decision nodes in $H_k(\tilde{s}_k(h_j^2))$. Since $s_k^*(h_j^2) \in S_k(h_k) \cap S_k(h_j^2)$, and h_k does not precede h_j^2, it follows that $\tilde{s}_k(h_j^2) \in S_k(h_k) \cap S_k(h_j^2)$ as well. Moreover, as $\tilde{s}_k(h_j^2)$ coincides with $s_k(h_j^2)$ in the subgame starting at h_k, we may conclude that $t_k(h_j^2)$ strictly prefers $\tilde{s}_k(h_j^2)$ over $s_k^*(h_j^2)$ at h_k. Since $t_j(h_i)$ believes at h_j^2 that player k is of type $t_k(h_j^2)$, the following holds:

$$t_j(h_i) \text{ believes at } h_j^2 \text{ that player } k, \text{ at } h_k, \text{ strictly prefers } \tilde{s}_k(h_j^2) \text{ over } s_k^*(h_j^2), \tag{5.1}$$

where both $\tilde{s}_k(h_j^2)$ and $s_k^*(h_j^2)$ are in $S_k(h_k) \cap S_k(h_j^2)$.

We have seen above that $t_j(h_i)$ satisfies minimal belief revision. Moreover, since every $t \in T^*$ satisfies Bayesian updating, we may conclude that $t_j(h_i)$ believes that every opponent satisfies Bayesian updating. By Theorem 4.9 we may thus conclude that $t_j(h_i)$ satisfies proper belief revision. Now, let h_j^1 be the unique decision node in H_j^* that precedes h_j^2 and for which no other player j decision node is between h_j^1 and h_j^2. Since both $\tilde{s}_k(h_j^2)$ and $s_k^*(h_j^2)$ are in $S_k(h_k) \cap S_k(h_j^2)$, proper belief revision of $t_j(h_i)$ together with (5.1) implies the following:

$$t_j(h_i) \text{ believes at } h_j^1 \text{ that player } k, \text{ at } h_k, \text{ strictly prefers } \tilde{s}_k(h_j^2) \text{ over } s_k^*(h_j^2). \tag{5.2}$$

As h_j^1 precedes h_j^2, and h_k does not precede h_j^2, we must have that h_k does not precede h_j^1. Hence, $s_k^*(h_j^1)$ prescribes at every player k decision node h_k' weakly following h_k the backward induction $a^*(h_k')$, just as $s_k^*(h_j^2)$ does. Together with (5.2), this yields:

$$t_j(h_i) \text{ believes at } h_j^1 \text{ that player } k, \text{ at } h_k, \text{ strictly prefers } \tilde{s}_k(h_j^2) \text{ over } s_k^*(h_j^1).$$

Since $t_j(h_i)$ believes at h_j^1 that player k has type $t_k(h_j^1)$, it follows that $s_k^*(h_j^1)$ is not sequentially rational for $t_k(h_j^1)$, which contradicts our induction assumption that $s_k^*(h_j^1)$ is sequentially rational for $t_k(h_j^1)$. Hence, we may conclude that $s_k^*(h_j^2)$ is optimal for $t_k(h_j^2)$ at every $h_k \in H_k(s_k^*(h_j^2))$ not preceding h_j^2. This completes Case 1.

Case 2. Assume that $h_k \in H_k(s_k^*(h_j^2))$ precedes h_j^2. Since $t_j(h_i)$ satisfies minimal belief revision, as we have seen above, it must be the case that the type $t_k(h_j^2) = t_k(t_j(h_i), h_j^2)$ rationalizes the event of reaching h_j^2. Hence, $t_k(h_j^2)$ has a sequentially rational strategy $s_k(h_j^2)$ in $S_k(h_j^2)$. Suppose, contrary to what we want to prove, that $s_k^*(h_j^2)$ is not optimal for $t_k(h_j^2)$ at h_k. Then, necessarily,

$$t_k(h_j^2) \text{ strictly prefers } z(s_k(h_j^2), t_k(h_j^2), h_k) \text{ over } z(s_k^*(h_j^2), t_k(h_j^2), h_k). \tag{5.3}$$

Since $s_k(h_j^2)$ and $s_k^*(h_j^2)$ are both in $S_k(h_j^2)$, they coincide on all player k decision nodes preceding h_j^2. Hence, by (5.3), there must be some player k decision node h_k' not preceding h_j^2 such that (1) $s_{-k}(t_k(h_j^2), h_k) \in S_{-k}(h_k')$, and (2) $(s_k(h_j^2), s_{-k}(t_k(h_j^2), h_k))$ and $(s_k^*(h_j^2), s_{-k}(t_k(h_j^2), h_k))$ both reach h_k'. By Bayesian updating of $t_k(h_j^2)$, we then have that $s_{-k}(t_k(h_j^2), h_k') = s_{-k}(t_k(h_j^2), h_k)$. This implies that

$$z(s_k(h_j^2), t_k(h_j^2), h_k) = z(s_k(h_j^2), t_k(h_j^2), h_k') \text{ and } z(s_k^*(h_j^2), t_k(h_j^2), h_k) = z(s_k^*(h_j^2), t_k(h_j^2), h_k').$$

Together with (5.3), we may conclude that

$$t_k(h_j^2) \text{ strictly prefers } z(s_k(h_j^2), t_k(h_j^2), h_k') \text{ over } z(s_k^*(h_j^2), t_k(h_j^2), h_k'),$$

which means that $s_k^*(h_j^2)$ is not optimal for $t_k(h_j^2)$ at h_k'. However, this contradicts our findings in Case 1, as h_k' does not precede h_j^2. Therefore, $s_k^*(h_j^2)$ must be optimal for $t_k(h_j^2)$ at h_k. This completes Case 2.

By combining the cases 1 and 2, we have shown for every $h_k \in H_k(s_k^*(h_j^2))$ that $s_k^*(h_j^2)$ is optimal for $t_k(h_j^2)$ at h_k. As such, $s_k^*(h_j^2)$ is sequentially rational for $t_k(h_j^2)$. Since $t_j(h_i)$ believes at h_j^2 that player k is of type $t_k(h_j^2)$ and chooses strategy $s_k^*(h_j^2)$, and since this holds for every h_j^2 and every opponent k, it follows that $t_j(h_i)$ believes in sequential rationality, which was to show.

We may thus conclude that every type $t \in T^*$ believes in sequential rationality, satisfies minimal belief revision and initially believes in P. As every type $t \in T^*$ believes that all opponents' types are in T^*, it holds that every type $t \in T^*$ respects common belief in the events that every type believes in sequential rationality, satisfies minimal belief revision and initially believes in P. This completes the proof of this theorem. ■

We now prove that common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in a profile P of preference relations, leads to backward induction in the game with preference relations P.

Theorem 5.2. *Let $\mathcal{S}$ be an extensive form structure with perfect information, and $P = (P_i)_{i \in I}$ a profile of preference relations on the set of terminal nodes. Let t_i be a type with preference relation P_i, respecting common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in P. Then, there is a unique sequentially rational strategy for t_i, namely player i's backward induction strategy in $(\mathcal{S}, P)$.*

Proof. For a given player i, decision node $h_i \in H_i^*$ and opponent j, let $S_j^*(h_i)$ be the set of player j strategies s_j such that (1) $s_j \in S_j(h_i)$, and (2) at every $h_j \in H_j(s_j)$ following h_i, the strategy s_j prescribes the backward induction action $a^*(h_j)$ in $(\mathcal{S}, P)$. We prove the following property.

Claim. Let t_i be a player i type that respects common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in P. Then,

$$s_j(t_i, h_i) \in S_j^*(h_i)$$

for all $h_i \in H_i^*$ and all opponents j.

Proof of Claim. We prove the claim by induction on the number of decision nodes following h_i. If h_i is not followed by any decision node, the statement is trivial since $S_j^*(h_i) = S_j(h_i)$. Suppose now that the claim holds for all pairs (i', j') of players and every decision node $h_{i'}$ followed by at most $K - 1$ decision nodes. Choose h_i with the property that h_i is followed by exactly K decision nodes. We prove that $s_j(t_i, h_i) \in S_j^*(h_i)$ for all opponents j. Hence, we must show that for every decision node $h_j \in H_j(s_j(t_i, h_i))$ following h_i, the strategy $s_j(t_i, h_i)$ prescribes the backward induction action $a^*(h_j)$.

Let $t_j^* = t_j(t_i, h_i)$ and $s_j^* = s_j(t_i, h_i)$. Choose a decision node $h_j \in H_j(s_j^*)$ following h_i. We shall prove that $s_j^*(h_j) = a^*(h_j)$. As t_i respects common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in P, and since t_i believes at h_i that player j is of type t_j^*, it follows that t_j^* respects common belief in the events that types believe in sequential rationality, satisfy minimal belief revision, and initially believe in P. Since h_j is followed by at most $K - 1$ decision nodes, we thus know by the induction assumption that

$$s_k(t_j^*, h_j) \in S_k^*(h_j)$$

for all opponents $k \neq j$. Consequently, t_j^* believes at h_j that all opponents choose their backward induction actions in $(\mathcal{S}, P)$ at the decision nodes following h_j.

As t_i initially believes in P, it follows that $t_j(t_i, h_0)$ has preference relation P_j. Moreover, since t_i satisfies minimal belief revision, it must be the case that $t_j(t_i, h_0)$ has the same conditional belief vector as $t_j(t_i, h_i) = t_j^*$. We may thus conclude that $t_j(t_i, h_0)$ believes at h_j that all opponents choose their backward induction actions in $(\mathcal{S}, P)$ at the decision nodes following h_j. Together with the fact that $t_j(t_i, h_0)$ has preference relation P_j, it follows that $t_j(t_i, h_0)$'s optimal strategies at h_j all prescribe the backward induction action $a^*(h_j)$ at h_j. More precisely, for every $s_j \in S_j(h_j)$ not prescribing $a^*(h_j)$ at h_j there is some $s_j' \in S_j(h_j)$ prescribing $a^*(h_j)$

at h_j such that $t_j(t_i, h_0)$ strictly prefers s'_j over s_j at h_j. This, in turn, means that t_i believes at h_0 that for every $s_j \in S_j(h_j)$ not prescribing $a^*(h_j)$ at h_j there is some $s'_j \in S_j(h_j)$ prescribing $a^*(h_j)$ at h_j such that player j strictly prefers s'_j over s_j at h_j.

Since t_i believes that all opponents believe in sequential rationality and satisfy minimal belief revision, we know by Lemma 4.4 that t_i believes that all opponents satisfy Bayesian updating. Together with the fact that t_i satisfies minimal belief revision, we may conclude by Theorem 4.9 that t_i satisfies proper belief revision. Therefore, t_i's belief at h_i about player j's preference relation at h_j over strategies in $S_j(h_j) \cap S_j(h_i)$ should coincide with t_i's belief at h_0 about player j's preference relation at h_j over strategies in $S_j(h_j) \cap S_j(h_i)$. Since, by assumption, h_j follows h_i we have that $S_j(h_j) \subseteq S_j(h_i)$. Hence, t_i's belief at h_i about player j's preference relation over strategies in $S_j(h_j)$ should coincide with t_i's belief at the beginning about player j's preference relation at h_j over strategies in $S_j(h_j)$. Since we have seen that t_i believes at h_0 that for every $s_j \in S_j(h_j)$ not prescribing $a^*(h_j)$ at h_j there is some $s'_j \in S_j(h_j)$ prescribing $a^*(h_j)$ at h_j such that player j strictly prefers s'_j over s_j at h_j, it follows that t_i believes so at h_i. This implies, however, that t_i believes at h_i that player j's optimal strategies at h_j all prescribe the backward induction action $a^*(h_j)$ at h_j.

Since t_i believes in sequential rationality, and since $s^*_j = s_j(t_i, h_i)$, we must have that s^*_j is optimal for $t_j(t_i, h_i)$ at h_j. By the above, it follows that s^*_j must prescribe the backward induction $a^*(h_j)$ at h_j, which was to show. This completes the proof of the claim.

Now, let t_i be a type that has preference relation P_i, and that respects common belief in the events that types believe in sequential rationality, satisfy minimal belief revision and initially believe in P. By the claim, we know that t_i believes at every decision node h_i that his opponents will choose the backward induction actions in $(\mathcal{S}, P)$ at every decision node following h_i. Since t_i has preference relation P_i, the unique sequentially rational strategy for t_i is his backward induction strategy in $(\mathcal{S}, P)$. This completes the proof. ■

6. Dropping Minimal Belief Revision

In the previous section we have seen that common belief in the events that types initially believe in P, believe in sequential rationality and satisfy minimal belief revision, singles out the backward induction strategy for player i in the game $(\mathcal{S}, P)$. In this section we investigate how crucial "minimal belief revision" is in establishing this relationship with backward induction. More precisely, we study the consequences of replacing the minimal belief revision requirement by the more basic condition of Bayesian updating. We shall prove that the resulting rationality concept allows for any strategy that survives the Dekel-Fudenberg procedure (Dekel and Fudenberg (1990)), that is, one round of elimination of weakly dominated strategies followed by iterative elimination of strongly dominated strategies. As to formally state this result, we must first define the Dekel-Fudenberg procedure in some more detail.

Let $\mathcal{S}$ be an extensive form structure with perfect information and $P = (P_i)_{i \in I}$ a profile of preference relations on the terminal nodes. For every player i, let u_i be an arbitrary utility

function on the terminal nodes that represents P_i, and let $u = (u_i)_{i \in I}$. Let $\Delta(S_i)$ be the set of probability distributions on the set S_i of player i strategies, and let $S_{-i} = \times_{j \neq i} S_j$ be the set of opponents' strategy profiles. For a pair $(\mu_i, s_{-i}) \in \Delta(S_i) \times S_{-i}$, we denote by

$$u_i(\mu_i, s_{-i}) = \sum\nolimits_{s_i \in S_i} \mu_i(s_i)\ u_i(z(s_i, s_{-i}))$$

the expected utility induced by (μ_i, s_{-i}) and the utility function u_i. Here, $z(s_i, s_{-i})$ denotes the terminal node reached by the strategy profile (s_i, s_{-i}). We say that strategy s_i is *weakly dominated* with respect to u_i if there is some $\mu_i \in \Delta(S_i)$ such that (1) $u_i(\mu_i, s_{-i}) \geq u_i(s_i, s_{-i})$ for all $s_{-i} \in S_{-i}$, and (2) $u_i(\mu_i, s_{-i}) > u_i(s_i, s_{-i})$ for some $s_{-i} \in S_{-i}$. Now, fix some subset $\tilde{S}_{-i} \subseteq S_{-i}$ of opponents' strategy profiles. We say that s_i is *strongly dominated on* $\tilde{S}_{-i}$ with respect to u_i if there is some $\mu_i \in \Delta(S_i)$ such that $u_i(\mu_i, s_{-i}) > u_i(s_i, s_{-i})$ for all $s_{-i} \in \tilde{S}_{-i}$.

For every player i, let $DF_i^1(u)$ be the set of strategies in S_i that are not weakly dominated with respect to u_i. For every player i and every $k \geq 2$, recursively define $DF_i^k(u)$ as the set of strategies in $DF_i^{k-1}(u)$ that are not strongly dominated on $\times_{j \neq i} DF_j^{k-1}(u)$ with respect to u_i. Finally, let $DF_i^\infty(u) = \cap_{k \in \mathbb{N}} DF_i^k(u)$ for every player i. We say that a strategy $s_i \in S_i$ survives the *Dekel-Fudenberg procedure* with respect to u if and only if $s_i \in DF_i^\infty(u)$. The following theorem states that replacing minimal belief revision by Bayesian updating in the model leads to the Dekel-Fudenberg procedure.

Theorem 6.1. *Let $\mathcal{S}$ be an extensive form structure with perfect information, $P = (P_i)_{i \in I}$ a profile of preference relations on the terminal nodes, and s_i a strategy for player i. Then, the following two statements are equivalent:*
(1) s_i is sequentially rational for some type t_i having preference relation P_i and respecting common belief in the events that types believe in sequential rationality, satisfy Bayesian updating, and initially believe in P;
(2) s_i survives the Dekel-Fudenberg procedure with respect to every u representing P.

Here, we say that u represents P if for every player i it holds that u_i represents P_i. In particular, the theorem implies that the Dekel-Fudenberg procedure for generic games with perfect information does not depend upon the particular utility functions that are chosen to represent the preference relations over terminal nodes. Since it is well-known that the Dekel-Fudenberg procedure may select strategies, and even outcomes, that are not compatible with backward induction, the minimal belief revision requirement may be seen as a property that closes the gap between the Dekel-Fudenberg procedure and the backward induction procedure.

In Ben-Porath (1997) it has been shown that also the concept of "common certainty of rationality at the beginning of the game" leads exactly to those strategies surviving the Dekel-Fudenberg procedure. The latter concept is, however, built upon fundamentally different principles than ours. Ben-Porath assumes, namely, that players believe *throughout the game* that the opponents hold preference relations as given by P, while we only assume players to believe so

at the beginning of the game. On the other hand, Ben-Porath only requires players to believe *at the beginning of the game* that his opponents choose sequentially rational strategies, while our "belief in sequential rationality" condition requires players to believe so *at each instance of the game*. One could therefore argue that the concept of "common certainty of rationality at the beginning of the game" is in some sense dual to the concept studied in this section. Nevertheless, both concepts eventually make the same selection of strategies for each player.

As a preparatory step towards proving Theorem 6.1, we first characterize weakly dominated and strongly dominated strategies in games with perfect information. These characterizations are based upon results in Ben-Porath (1997) and Pearce (1984), and are stated in the following lemma.[3] In this lemma, we say that a strategy s_i is *initially rational* for type t_i if s_i is optimal for t_i against the initial belief $s_{-i}(t_i, h_0)$.

Lemma 6.2. *Let $\mathcal{S}$ be an extensive form structure with perfect information, P_i a preference relation on the terminal nodes, u_i a utility function representing P_i, and s_i a strategy for player i. Then the following is true:*
(1) s_i is not weakly dominated with respect to u_i if and only if s_i is sequentially rational for some type t_i that has preference relation P_i and satisfies Bayesian updating;
(2) s_i is not strongly dominated on some $\tilde{S}_{-i} \subseteq S_{-i}$ with respect to u_i if and only if s_i is initially rational for some type t_i with preference relation P_i and initial belief $s_{-i}(t_i, h_0) \in \tilde{S}_{-i}$.

Proof. (1) Suppose that s_i is not weakly dominated with respect to u_i. By Lemma 4 in Pearce (1984), there is some $\mu_{-i} \in \Delta(S_{-i})$ with full support such that s_i is optimal against μ_{-i} with respect to u_i. Then, by Lemma 1.1 in Ben-Porath (1997), there exists a *probabilistic* updating system $(\mu_{-i}(h_i))_{h_i \in H_i}$ with $\mu_{-i}(h_i) \in \Delta(S_{-i}(h_i))$ for every $h_i \in H_i$ such that (a) this updating system satisfies Bayesian updating, and (b) for every $h_i \in H_i(s_i)$, strategy s_i is optimal against $\mu_{-i}(h_i)$ with respect to u_i. By Lemma 1.2.1 in Ben-Porath (1997), there then exists a *deterministic* updating system $(s_{-i}(h_i))_{h_i \in H_i}$ with $s_{-i}(h_i) \in S_{-i}(h_i)$ for every $h_i \in H_i$ such that (a) this updating system satisfies Bayesian updating, and (b) for every $h_i \in H_i(s_i)$, strategy s_i is optimal against $s_{-i}(h_i)$ with respect to u_i. One can then choose a type t_i satisfying Bayesian updating, with preference relation P_i, and with $s_{-i}(t_i, h_i) = s_{-i}(h_i)$ for all $h_i \in H_i$. Hence, by construction, s_i is sequentially rational for t_i.

Now, suppose that s_i is sequentially rational for some type t_i that has preference relation P_i and satisfies Bayesian updating. Hence, for every $h_i \in H_i(s_i)$, the strategy s_i is optimal against $s_{-i}(t_i, h_i) \in S_{-i}(h_i)$ with respect to u_i. By Lemma 1.2 in Ben-Porath (1997), it then follows that s_i is optimal with respect to u_i against some $\mu_{-i} \in \Delta(S_{-i})$ with full support. Lemma 4 in Pearce (1984) implies that s_i is not weakly dominated with respect to u_i.

(2) Suppose that s_i is not strongly dominated on some $\tilde{S}_{-i} \subseteq S_{-i}$ with respect to u_i. By Lemma 3 in Pearce (1984), we then know that there is some $\mu_{-i} \in \Delta(\tilde{S}_{-i})$ such that s_i is

[3] Formally, Lemmas 3 and 4 in Pearce (1984), which are the results we use here, are stated for two-player games only. However, if one allows for *correlated* probability distributions on the opponents' strategy spaces, as we do, then Pearce's results also apply to more than two players.

optimal against μ_{-i} with respect to u_i. By the proof of Lemma 1.2.1 in Ben-Porath (1997), we may conclude that there is some s_{-i} in the support of μ_{-i} such that s_i is optimal against s_{-i} with respect to u_i (and hence with respect to P_i). Let t_i be a type with preference relation P_i and $s_{-i}(t_i, h_0) = s_{-i}$. Then, $s_{-i}(t_i, h_0) \in \tilde{S}_{-i}$ and s_i is initially rational for t_i.

Assume finally that s_i is initially rational for some type t_i with preference relation P_i and $s_{-i}(t_i, h_0) \in \tilde{S}_{-i}$. Then, s_i is optimal against $s_{-i}(t_i, h_0) \in \tilde{S}_{-i}$ with respect to u_i. But then, s_i is obviously not strongly dominated on $\tilde{S}_{-i}$ with respect to u_i. This completes the proof of this lemma. ■

By means of this lemma, we are now able to provide an alternative characterization of strategies that survive the Dekel-Fudenberg procedure. Let $P = (P_i)_{i \in I}$ be a profile of preference relations. For every player i, let $S_i^1(P)$ be the set of strategies that are sequentially rational for some type t_i with preference relation P_i that satisfies Bayesian updating. For $k \geq 2$, recursively define $S_i^k(P)$ as the set of strategies in $S_i^{k-1}(P)$ that are sequentially rational for some t_i with preference relation P_i, satisfying Bayesian updating, and with initial belief $s_{-i}(t_i, h_0)$ in $\times_{j \neq i} S_j^{k-1}(P)$. Finally, let $S_i^\infty(P) = \cap_{k \in \mathbb{N}} S_i^k(P)$. We obtain the following characterization.

Lemma 6.3. *Let $\mathcal{S}$ be an extensive form structure with perfect information and $P = (P_i)_{i \in I}$ a profile of preference relations. Then, $S_i^\infty(P) = DF_i^\infty(u)$ for every profile u of utility functions representing P.*

Proof. Choose a profile u of utility functions representing P. We show, by induction on k, that $S_i^k(P) = DF_i^k(u)$ for all players i. By Lemma 6.2, part (1), we know that $S_i^1(P) = DF_i^1(u)$. Now, let $k \geq 2$ and assume that $S_j^{k-1}(P) = DF_j^{k-1}(u)$ for all players j. We first show that $S_i^k(P) \subseteq DF_i^k(u)$. Take some arbitrary $s_i \in S_i^k(P)$. Hence, s_i is sequentially rational for some type t_i with preference relation P_i, satisfying Bayesian updating, and with $s_{-i}(t_i, h_0) \in \times_{j \neq i} S_j^{k-1}(P)$. Since t_i satisfies Bayesian updating, the fact that s_i is sequentially rational for t_i implies that s_i is initially rational for t_i. By Lemma 6.2, part (2), it follows that s_i is not strongly dominated on $\times_{j \neq i} S_j^{k-1}(P)$ with respect to u_i. Since, by induction assumption, $S_j^{k-1}(P) = DF_j^{k-1}(u)$ for all $j \neq i$, it follows that s_i is not strongly dominated on $\times_{j \neq i} DF_j^{k-1}(u)$ with respect to u_i. On the other hand, we know that $s_i \in S_i^k(P) \subseteq S_i^{k-1}(P)$ which, by induction assumption, is equal to $DF_i^{k-1}(u)$. Hence, $s_i \in DF_i^{k-1}(u)$ and s_i is not strongly dominated on $\times_{j \neq i} DF_j^{k-1}(u)$ with respect to u_i, which implies that $s_i \in DF_i^k(u)$.

We finally show that $DF_i^k(u) \subseteq S_i^k(P)$. Take some arbitrary $s_i \in DF_i^k(u)$. Hence, s_i is not strongly dominated on $\times_{j \neq i} DF_j^{k-1}(u)$ with respect to u_i. By Lemma 6.2, part (2), we then have that s_i is initially rational for some type t_i with preference relation P_i and $s_{-i}(t_i, h_0) \in \times_{j \neq i} DF_j^{k-1}(u)$. Since, by induction assumption, $DF_j^{k-1}(u) = S_j^{k-1}(P)$ for all $j \neq i$, we thus have that s_i is initially rational for some type t_i with preference relation P_i and $s_{-i}(t_i, h_0) \in \times_{j \neq i} S_j^{k-1}(P)$. As $s_i \in DF_i^k(u) \subseteq DF_i^1(u)$ and, by induction assumption, $DF_i^1(u) = S_i^1(P)$, it follows that $s_i \in S_i^1(P)$. Hence, there is some type t_i' with preference relation P_i and satisfying

Bayesian updating, such that s_i is sequentially rational for t_i'. Now, construct a type t_i'' with the following properties: (1) t_i'' has preference relation P_i, (2) $s_{-i}(t_i'', h_i) = s_{-i}(t_i, h_0)$ at all $h_i \in H_i^*$ for which $s_{-i}(t_i, h_0) \in S_{-i}(h_i)$, and (3) $s_{-i}(t_i'', h_i) = s_{-i}(t_i', h_i)$ at all other $h_i \in H_i^*$. Then, by construction, s_i is sequentially rational for t_i''. Since, moreover, t_i'' has preference relation P_i, satisfies Bayesian updating and has initial belief $s_{-i}(t_i'', h_0) = s_{-i}(t_i, h_0) \in \times_{j \neq i} S_j^{k-1}(P)$, we thus have that s_i is sequentially rational for a type with preference relation P_i, satisfying Bayesian updating, and with initial belief in $\times_{j \neq i} S_j^{k-1}(P)$. On the other hand, $s_i \in DF_i^k(u) \subseteq DF_i^{k-1}(u)$ which, by induction assumption, is equal to $S_i^{k-1}(P)$. Hence, $s_i \in S_i^{k-1}(P)$. Together with the previous insight, we may thus conclude that $s_i \in S_i^k(P)$. It thus follows that $S_i^k(P) = DF_i^k(u)$ for all players i and all k, which implies that $S_i^\infty(P) = DF_i^\infty(u)$ for all players i. This completes the proof. ■

We are now in a position to prove Theorem 6.1.

Proof of Theorem 6.1. For every player i, let $T_i^*(P)$ be the set of player i types that have preference relation P_i and respect common belief in the events that types believe in sequential rationality, satisfy Bayesian updating and initially believe in P. Let $S_i^*(P)$ be the set of player i strategies that are sequentially rational for some type in $T_i^*(P)$.

We first show the implication from (1) to (2). Let u be an arbitrary profile of utility functions that represents P. We show that $S_i^*(P) \subseteq DF_i^\infty(u)$. By Lemma 6.3 we know that $DF_i^\infty(u) = S_i^\infty(P)$, and hence it is sufficient to show that $S_i^*(P) \subseteq S_i^\infty(P)$, which in turn is equivalent to showing that $S_i^*(P) \subseteq S_i^k(P)$ for every k. We prove the latter claim by induction on k.

For $k = 1$, we must show that $S_i^*(P) \subseteq S_i^1(P)$. Let $s_i \in S_i^*(P)$. Then, by definition of $S_i^*(P)$, s_i is sequentially rational for some type t_i satisfying Bayesian updating and having preference relation P_i, and hence $s_i \in S_i^1(P)$. We may thus conclude that $S_i^*(P) \subseteq S_i^1(P)$ for all players i.

Now, let $k \geq 2$, and assume that $S_j^*(P) \subseteq S_j^{k-1}(P)$ for every player j. Choose an arbitrary player i. We prove that $S_i^*(P) \subseteq S_i^k(P)$. Choose some $s_i \in S_i^*(P)$. Then, there is some type t_i with preference relation P_i and respecting common belief in the events that types believe in sequential rationality, satisfy Bayesian updating and initially believe in P, such that s_i is sequentially rational for t_i. Fix an opponent j. Then, it follows that strategy $s_j(t_i, h_0)$ is sequentially rational for type $t_j(t_i, h_0)$, and, moreover, type $t_j(t_i, h_0)$ has preference relation P_j and respects common belief in the events that types believe in sequential rationality, satisfy Bayesian updating and initially believe in P. Hence, $t_j(t_i, h_0) \in T_j^*(P)$. Since $s_j(t_i, h_0)$ is sequentially rational for $t_j(t_i, h_0) \in T_j^*(P)$, it follows that $s_j(t_i, h_0) \in S_j^*(P)$. We may thus conclude that $s_j(t_i, h_0) \in S_j^*(P)$ for every opponent j and hence, by the induction assumption, $s_j(t_i, h_0) \in S_j^{k-1}(P)$ for all opponents j. Therefore, s_i is sequentially rational for a type t_i that has preference relation P_i, satisfies Bayesian updating, and has initial belief $s_{-i}(t_i, h_0)$ in $\times_{j \neq i} S_j^{k-1}(P)$. On the other hand we know that $s_i \in S_i^*(P)$ which, by induction assumption, is a subset of $S_i^{k-1}(P)$. It thus follows that $s_i \in S_i^k(P)$. By induction, we may thus conclude that

$S_i^*(P) \subseteq S_i^k(P)$ for every k and every player i, and hence $S_i^*(P) \subseteq S_i^\infty(P)$ for every player i. Hence, $S_i^*(P) \subseteq DF_i^\infty(u)$ for every player i. The implication from (1) and (2) thus follows.

We now show the implication from (2) to (1). Let u be a profile of utility functions that represents P. We must show that $DF_i^\infty(P) \subseteq S_i^*(P)$ for all players i. By Lemma 6.3 we know that $DF_i^\infty(P) = S_i^\infty(P)$, and hence it is sufficient to show that $S_i^\infty(P) \subseteq S_i^*(P)$ for every player i.

By construction of $S_i^\infty(P)$, we may find for every $s_i \in S_i^\infty(P)$ some type t_i with preference relation P_i, satisfying Bayesian updating and having initial belief $s_{-i}(t_i, h_0)$ in $\times_{j \neq i} S_j^\infty(P)$ such that s_i is sequentially rational for t_i. Hence, for every $s_i \in S_i^\infty(P)$ there is an updating system $c_i(s_i) = (c_i(s_i)(h_i))_{h_i \in H_i^*}$ with $c_i(s_i)(h_i) \in S_{-i}(h_i)$ for every $h_i \in H_i^*$ (see Section 4.1) and $c_i(s_i)(h_0) \in \times_{j \neq i} S_j^\infty(P)$, such that the updating system satisfies Bayesian updating, and s_i is sequentially rational with respect to $c_i(s_i)$ and P_i.

For every $s_i \notin S_i^\infty(P)$ we may find some updating system $c_i(s_i)$ with $c_i(s_i)(h_0) \in \times_{j \neq i} S_j^\infty(P)$, and preference relation $P_i(s_i)$, not necessarily equal to P_i, such that $c_i(s_i)$ satisfies Bayesian updating and s_i is sequentially rational with respect to $c_i(s_i)$ and $P_i(s_i)$. For every $s_i \in S_i^\infty(P)$, simply set $P_i(s_i) = P_i$.

Now, suppose that these updating systems $c_i(s_i)$ and preference relations $P_i(s_i)$ have been defined for all players i and strategies s_i. We may construct for every player i and every strategy s_i a type $t_i(s_i)$ with the following properties: (a) the preference relation of $t_i(s_i)$ is given by $P_i(s_i)$, (b) for every $h_i \in H_i^*$ and opponent j, the conditional belief $s_j(t_i(s_i), h_i)$ about player j's strategy choice is given by $c_{ij}(s_i)(h_i)$, where $c_{ij}(s_i)(h_i)$ is the belief at h_i about player j's strategy choice in the updating system $c_i(s_i)$, and (c) for every $h_i \in H_i^*$ and opponent j, the conditional belief $t_j(t_i(s_i), h_i)$ about player j's type is given by $t_j(s_j)$, where $s_j = c_{ij}(s_i)(h_i)$.

Claim. Every type $t_i(s_i)$ respects common belief in the events that types believe in sequential rationality, satisfy Bayesian updating and initially believe in P.

Proof of claim. Define $\tilde{T} = \{t_i(s_i) \mid i \in I \text{ and } s_i \in S_i\}$. By construction of the types $t_i(s_i)$, we have that $t_j(t_i(s_i), h_i) \in \tilde{T}$ for every player i, type $t_i(s_i) \in \tilde{T}$, decision node $h_i \in H_i^*$ and opponent j. Hence, in order to show the claim, it is sufficient to show that every type in $\tilde{T}$ believes in sequential rationality, satisfies Bayesian updating and initially believes in P.

Belief in sequential rationality. Let $t_i(s_i)$ be a type in $\tilde{T}$. At every $h_i \in H_i^*$, the type $t_i(s_i)$ believes that opponent j chooses strategy $s_j = c_{ij}(s_i)(h_i)$ and believes that opponent j has type $t_j(s_j)$. By construction, type $t_j(s_j)$'s conditional belief about the opponents' strategy choices is given by $c_j(s_j)$. Since s_j is sequentially rational for $c_j(s_j)$, it follows that s_j is sequentially rational for $t_j(s_j)$. It therefore follows that the strategy $s_j(t_i(s_i), h_i) = s_j$ is sequentially for $t_j(t_i(s_i), h_i) = t_j(s_j)$ for every $h_i \in H_i^*$ and every opponent j, which implies that $t_i(s_i)$ believes in sequential rationality.

Bayesian updating. Bayesian updating of $t_i(s_i) \in \tilde{T}$ follows immediately from the fact that $t_i(s_i)$'s conditional beliefs about the opponents' strategy choices is given by $c_i(s_i)$, and the assumption that $c_i(s_i)$ satisfies Bayesian updating.

Initial belief in P. Let $t_i(s_i)$ be a type in $\tilde{T}$. Fix an opponent j. By definition, we have

that $s_j(t_i(s_i), h_0) = c_{ij}(s_i)(h_0)$. Since, by construction, $c_i(s_i)(h_0) \in \times_{j \neq i} S_j^{\infty}(P)$, we have that $c_{ij}(s_i)(h_0) \in S_j^{\infty}(P)$. Hence, $t_i(s_i)$ initially believes that player j chooses some strategy $s_j \in S_j^{\infty}(P)$. As such, $t_i(s_i)$ initially believes that player j has type $t_j(s_j)$, which has preference relation $P_j(s_j) = P_j$, since $s_j \in S_j^{\infty}$. We may thus conclude that $t_i(s_i)$ initially believes that player j has preference relation P_j. Since this holds for all opponents j, it follows that $t_i(s_i)$ initially believes in P.

We have thus shown that every type $t_i(s_i)$ in $\tilde{T}$ believes in sequential rationality, satisfies Bayesian updating, and initially believes in P. This implies the statement in the claim.

Recall that it is our objective to show that $S_i^{\infty}(P) \subseteq S_i^*(P)$. Take some arbitrary $s_i \in S_i^{\infty}(P)$. Then, s_i is sequentially rational for the type $t_i(s_i)$, and $t_i(s_i)$ has preference relation $P_i(s_i) = P_i$. By the claim above, it follows that $t_i(s_i)$ has preference relation P_i and respects common belief in the events that types believe in sequential rationality, satisfy Bayesian updating and initially believe in P. Hence, $t_i(s_i) \in T_i^*(P)$. Since s_i is sequentially rational for t_i, we have that $s_i \in S_i^*(P)$. We thus have shown that $S_i^{\infty}(P) \subseteq S_i^*(P)$, which implies that $DF_i^{\infty}(u) \subseteq S_i^*(P)$ for all players i. This establishes the implication from (2) to (1), and completes the proof of this theorem. ■

7. Appendix

Proof of Lemma 4.8. Let u^1 be an arbitrary utility representation of P^1, and let the utility functions u^2 and $\tilde{u}^2$ be as stated in the lemma. Let $D(P^1, P^2)$ be the set of unordered pairs of terminal nodes on which P^1 and P^2 disagree. Similarly, we define $D(P^1, \tilde{P}^2)$. Without loss of generality, let a and b in the lemma be chosen such that $u^1(a) > u^1(b)$. Then, by construction, $u^2(a) < u^2(b)$ and $\tilde{u}^2(a) > \tilde{u}^2(b)$. We prove our result through a series of smaller facts. The proof for each of these facts is given in the lines immediately following the statement of the fact.
Fact 1. It holds that $\{a, b\} \notin D(P^1, \tilde{P}^2)$, but $\{a, b\} \in D(P^1, P^2)$.
This follows directly from the observation that $u^1(a) > u^1(b)$, $\tilde{u}^2(a) > \tilde{u}^2(b)$ but $u^2(a) < u^2(b)$.
Fact 2. Let $\{x, y\} \in D(P^1, \tilde{P}^2)$, and $x, y \notin \{a, b\}$. Then, $\{x, y\} \in D(P^1, P^2)$.
This follows directly from the observation that $\tilde{u}^2(x) = u^2(x)$ and $\tilde{u}^2(y) = u^2(y)$.
Fact 3. Let $\{a, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(y) > \tilde{u}^2(a)$. Then, $\{a, y\} \in D(P^1, P^2)$.
Since $\{a, y\} \in D(P^1, \tilde{P}^2)$ and $\tilde{u}^2(a) < \tilde{u}^2(y)$, we must have that $u^1(a) > u^1(y)$. On the other hand, by construction of $\tilde{u}^2$, we know that $u^2(a) = \tilde{u}^2(b)$ and $u^2(y) = \tilde{u}^2(y)$. Since $\tilde{u}^2(y) > \tilde{u}^2(a)$ and $\tilde{u}^2(a) > \tilde{u}^2(b)$, it follows that $u^2(a) = \tilde{u}^2(b) < \tilde{u}^2(y) = u^2(y)$, which implies that $\{a, y\} \in D(P^1, P^2)$.
Fact 4. Let $\{a, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(a) > \tilde{u}^2(y) > \tilde{u}^2(b)$. Then, $\{b, y\} \in D(P^1, P^2)$.
Since $\{a, y\} \in D(P^1, \tilde{P}^2)$ and $\tilde{u}^2(a) > \tilde{u}^2(y)$, we must have that $u^1(a) < u^1(y)$. By assumption, $u^1(a) > u^1(b)$, and hence $u^1(b) < u^1(y)$. By definition of $\tilde{u}^2$, we have that $u^2(b) = \tilde{u}^2(a)$ and $u^2(y) = \tilde{u}^2(y)$. Since $\tilde{u}^2(a) > \tilde{u}^2(y)$, we have that $u^2(b) > u^2(y)$, which implies that $\{b, y\} \in D(P^1, P^2)$.
Fact 5. Let $\{a, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(y) < \tilde{u}^2(b)$. Then, $\{a, y\} \in D(P^1, P^2)$.

As $\tilde{u}^2(y) < \tilde{u}^2(b)$ and $\tilde{u}^2(a) > \tilde{u}^2(b)$, we may conclude that $\tilde{u}^2(a) > \tilde{u}^2(y)$. Since $\{a, y\} \in D(P^1, \tilde{P}^2)$ we must have that $u^1(a) < u^1(y)$. By definition of $\tilde{u}^2$, it is seen that $u^2(y) = \tilde{u}^2(y)$ and $u^2(a) = \tilde{u}^2(b)$. As $\tilde{u}^2(b) > \tilde{u}^2(y)$, it follows that $u^2(a) > u^2(y)$, and hence $\{a, y\} \in D(P^1, P^2)$.
Fact 6. Let $\{b, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(y) > \tilde{u}^2(a)$. Then, $\{b, y\} \in D(P^1, P^2)$.
As $\tilde{u}^2(b) < \tilde{u}^2(a) < \tilde{u}^2(y)$, and $\{b, y\} \in D(P^1, \tilde{P}^2)$, we must have that $u^1(b) > u^1(y)$. By definition of $\tilde{u}^2$, it holds that $u^2(b) = \tilde{u}^2(a)$ and $u^2(y) = \tilde{u}^2(y)$. Since $\tilde{u}^2(a) < \tilde{u}^2(y)$, we know that $u^2(b) < u^2(y)$, and hence $\{b, y\} \in D(P^1, P^2)$.
Fact 7. Let $\{b, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(a) > \tilde{u}^2(y) > \tilde{u}^2(b)$. Then, $\{a, y\} \in D(P^1, P^2)$.
As $\tilde{u}^2(b) < \tilde{u}^2(y)$ and $\{b, y\} \in D(P^1, \tilde{P}^2)$, we may conclude that $u^1(b) > u^1(y)$. Since $u^1(a) > u^1(b)$, it follows that $u^1(a) > u^1(y)$. On the other hand, we know by definition of $\tilde{u}^2$ that $u^2(a) = \tilde{u}^2(b)$ and $u^2(y) = \tilde{u}^2(y)$. As $\tilde{u}^2(b) < \tilde{u}^2(y)$, it follows that $u^2(a) < u^2(y)$, and hence $\{a, y\} \in D(P^1, P^2)$.
Fact 8. Let $\{b, y\} \in D(P^1, \tilde{P}^2)$ such that $\tilde{u}^2(y) < \tilde{u}^2(b)$. Then, $\{b, y\} \in D(P^1, P^2)$.
Since $\tilde{u}^2(b) > \tilde{u}^2(y)$ and $\{b, y\} \in D(P^1, \tilde{P}^2)$, it must be the case that $u^1(b) < u^1(y)$. By construction of $\tilde{u}^2$, it holds that $u^2(b) = \tilde{u}^2(a)$ and $u^2(y) = \tilde{u}^2(y)$. As $\tilde{u}^2(a) > \tilde{u}^2(b) > \tilde{u}^2(y)$, we have that $u^2(b) > u^2(y)$, and hence $\{b, y\} \in D(P^1, P^2)$.

From Facts 1 to 8, it follows that $D(P^1, \tilde{P}^2)$ contains strictly less pairs than $D(P^1, P^2)$, and hence $d(P^1, \tilde{P}^2) < d(P^1, P^2)$. This completes the proof. ■

References

[1] Asheim, G.B. (2002), On the epistemic foundation for backward induction, *Mathematical Social Sciences* **44,** 121-144.

[2] Asheim, G.B. and A. Perea (2004), Sequential and quasi-perfect rationalizability in extensive games, Forthcoming in *Games and Economic Behavior.*

[3] Aumann, R. (1995), Backward induction and common knowledge of rationality, *Games and Economic Behavior* **8,** 6-19.

[4] Balkenborg, D. and E. Winter (1997), A necessary and sufficient epistemic condition for playing backward induction, *Journal of Mathematical Economics* **27,** 325-345.

[5] Battigalli, P. and M. Siniscalchi (1999), Hierarchies of conditional beliefs, and interactive epistemology in dynamic games, *Journal of Economic Theory* **88,** 188-230.

[6] Ben-Porath, E. (1997), Rationality, Nash equilibrium and backwards induction in perfect-information games, *Review of Economic Studies* **64,** 23-46.

[7] Dekel, E. and Fudenberg, D. (1990), Rational behavior with payoff uncertainty, *Journal of Economic Theory* **52,** 243-267.

[8] Ha, V. and Haddawy, P. (1998), Towards case-based preference elicitation: Similarity measures on preference structures, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence,* 193-201.

[9] Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52,** 1029-1050.

[10] Perea, A. (2002), A note on the one-deviation property in extensive form games, *Games and Economic Behavior* **40,** 322-338.

[11] Perea, A. (2003a), Forward induction and the minimum revision principle, Maastricht University.

[12] Perea, A. (2003b), Rationalizability and minimal complexity in dynamic games, Maastricht University.

[13] Perea, A. (2004), Proper rationalizability and belief revision in dynamic games, Maastricht University.

[14] Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* **59,** 909-924.

[15] Samet, D. (1996), Hypothetical knowledge and games with perfect information, *Games and Economic Behavior* **17,** 230-251.

[16] Schulte, O. (2002), Minimal belief change, Pareto-optimality and logical consequence, *Economic Theory* **19,** 105-144.

[17] Stalnaker, R. (1998), Belief revision in games: forward and backward induction, *Mathematical Social Sciences* **36,** 31-56.